확장한 어휘적 중의성 제거 규칙에 따른 부분 문장 분석에 기반한 한국어 문법 검사기

박 수 호 ∪ 권 혁 철
부 산 대 학 교 전 자 계 산 학 과
cs9334@hanmail.net, hckwon@hyowon.cc.pusan.ac.kr

A Korean Grammar Chacker Founded on Expanded Lexical Disambiguation Rule and Partial Parsing

Su-Ho Park ∪ Hyuk-chul Kwon
Dept. of Computer Science, Pusan Nation University

요 약

본 논문에서는 한국어 형태소 분석기가 처리할 수 없는 어휘적 중의성 해결을 위한 방법으로 부분 문장 분석 기법을 연구한다. 부분 문장 분석 기법의 신뢰도를 높이기 위해서 말뭉치를 이용한 데이터를 통해 학습한 경험적 규칙을 이용한다. 학습한 경험적 규칙을 오류 유형에 따라 확장하고 전문화하여 축적된 연구결과를 지식 베이스로 삼아 한국어 맞춤법 및 문법 검사기에서 사용하는 부분 문장 분석기의 성능을 향상시킨다. 본 논문에서 사용한 확장하고 전문화한 지식 베이스는 말뭉치에서 얻은 경험적 규칙을 기반으로 한다. 이 경험적 규칙은 언어적 지식을 기반으로 한다.

1. 서론

1.1 연구배경

가장 이상적인 맞춤법 및 문법 검사기는 문서 내에서 다수 어절에 걸쳐 생기는 맞춤법 및 문법 오류를 정확히 검증하고 대치어를 제시하는 것이다. 그러나 이것은 이상적인 맞춤법 및 문법 검사기일 뿐이고, 실제는 그렇지 못하다. 기존의 맞춤법 검사기는 한 어절을 대상으로 분석하기 때문에 문맥상 어울리지 않는 단어는 제대로 처리하지 못하였다. 문맥상 어울리지 않은 단어를 처리하기 위해서 구문 분석을 행해야 한다. 구문 분석을 하면 다수 어절에서 발생하는 구문 오류를 정확히 찾을 수는 있으나, 맞춤법 검사기의 속도를 떨어뜨리고 시스템 구현에서 많은 기억 공간을 요구하여 효율적이지 않다. 이런 문제를 해결하려고 부산대학교에서 개발한 맞춤법 및 문법 검사기는 맞춤법 검사기의 속도와 구문 오류를 고려하여 부분 문장 분석 기법을 사용한다. 부분 문장 분석 기법은 한국어 문서에 자주 나타나는 오류를 유형별로 분류하여 의존 관계를 바탕으로 지식베이스화했다. 또한 본 논문에서는 유형별로 나누어진 오류 유형별 지식 베이스를 확장하여 맞춤법 및 문법 검사기의 성능을 개선시켰다.

1.2 지식베이스와 부문 문장 분석 기법

다음[표-1]은 현재 문법 검사기에서 처리하는 오류 종류와 예제이다[6].

의미 오류 유형	예제 문장
적절하지 못한 단어의 사용	영희는 매우 공부하였다(X). 영희는 열심히 공부하였다(O).
발음상의 오류	더 나은 시스템을 계발하였다(X). 더 나은 시스템을 개발하였다(O).
철자의 오류	책이 책보에 쌓여 있다(X). 책이 책보에 싸여 있다(O).
의미상의 오류	그 형제는 우애가 두껍다(X). 그 형제는 우애가 두텁다(O).
문법적인 오류	알맞는 답을 고르시오(X). 알맞은 답을 고르시오(O).
복합명사 결합 오류	공부일생(X). 어머니일생(O).

[표-1] 오류 종류 및 예시

[표-1]과 같은 의미 오류 유형을 중심으로 구성된 지식베이스화한 규칙을 이용하여 구문 오류를 처리하는 것이 부문 문장 분석 기법이다. 부분 문장 분석 기법은 문장에서 오류를 일으킬 수 있는 검사 단어를 기준으로 의존 문법을 이용하여 다수 어절에서 연어 오류 단어를 찾아 교정한다.

부분 문장 분석을 하는 가장 큰 이유는 형태소 분석기로는 해결하지 못하는 형태소 분석만으로 해결할 수 없는 어휘나 의미의 중의성을 제거하기 위함이다. 부분 문장 분석기가 중의성이 발생한 의미 오류와 문체 오류를 찾는 과정에서 다음과 같은 문제를 일으킨다.

가) 복합 문장을 분석할 때 검사 단어가 속한 문장의 범위를 벗어나 연어 오류 단어를 찾는다.

나) 연어 오류 단어의 위치를 찾아내지 못하고 부분 문장 분석을 끝낸다.

가)와 나)의 문제 때문에 부분 문장 분석기가 검증을 제대로 할 수 없다. 다수 어절 내에서 어휘적 중의성이란 한 단어가 좌우 문맥을 고려하지 않은 상태에서 두 가지 이상의 형태소 정보를 가지는 현상을 말한다. 따라서 어휘적 중의성을 제거하기 위해 좌우 문맥을 고려하여 형태소 분석기로부터 불필요한 형태소 정보를 없애고 필요한 정보만을 가져와 검증하는 것이 부분 문장 분석 기법이다. 본 논문에서는 어휘적 중의성 때문에 발생하는 문제점을 제시하고 해결 방법을 제안하며, 부분 문장 분석기에서 쓰는 지식 베이스를 확장하여 부분 문장 분석기의 정확도를 높였다.

2. 어휘 중의성에 의해 발생하는 오류 유형과 문법 검사기

맞춤법 검사기는 현재 어절의 오류 여부만 판단하는 시스템이다. 따라서 어절과 어절 사이에서 발생하는 어휘적 중의성을 고려하지 않아도 상관없다. 그러나 문법 검사기는 다르다. 현재 어절의 오류 여부만을 판단해서는 어휘적 중의성이 발생할 때 이를 처리할 수 없다. 검사 단어를 중심으로 연어 오류 단어를 찾는 과정에서 부분 문장 분석을 해야 하므로 어휘적 중의성을 제거하지 않고는 정확한 검사가 불가능하다. 다음은 문법 검사기에서 사용하는 부분 문장 분석 기법과 오류 유형에 관한 것이다.

2.1 의존 문법을 이용한 부분 문장 분석

문장 분석기는 크게 어절 버퍼 관리자, 어휘 분석 모듈, 중의성 제거 모듈, 의존 관계 결정 모듈로 구분된다. 어절 버퍼 관리자는 어휘적 중의성 제거를 위해 현재 어절에서 분석 방향에 따라 앞 어절이나 뒤 어절을 가져오고, 어휘 분석 모듈은 어절 버퍼 관리자가 가져온 어절과 현재 어절의 형태소 분석 정보를 이용해서 의존 관계를 결정할 때 필요한 세부 정보를 저장한다. 그리고 의존 관계 결정 모듈은 의존 규칙에 따라 다수 어절 내에서 지배소와 피지배소간의 의존 관계를 조사한다. 의존 문법에서는 의존 가능한 두 가지의 어절 유형을 규정한 것이 문법 규칙이며[2], 의존 문법에서 의존관계의 주가 되는 것이 지배소이며, 지배소에 의한 규칙의 지배를 받는 것이 피지배소이다[7].

본 논문에서는 의존 관계를 결정할 때 의존 관계가 성립하는 규칙보다는 의존 관계가 성립하지 않는 규칙을 이용해서 의존 관계를 결정한다. 이 방법은 의존 관계 허용 범위를 넓혀 연어 오류 단어를 찾을 수 있는 범위를 확대하고, 의존 관계를 설정할 때 소요되는 시간을 줄인다. 그러나 허용 범위가 너무 확대되어 생기는 문법 검사기의 속도 저하와 검사 오류를 피할 수 없다. 이 문제를 해결하려고 문법 검사기가 검증할 수 있는 어절 수를 제한하거나 문장 분석을 종료하는 조건(Stop-Condition)을 둔다.

문장 분석 종료 조건(Stop-Condition)은 현재 어절을 분석할 때 용언의 종결형, 문장 종결 부호(!,?,.,)이 올 때이다. 부분 문장 분석기가 처리하는 검사 단어와 연어 오류 단어는 대체로 같은 주어나 같은 서술어를 가지는 문장 범위 안에 존재한다. 이는 문법 검사기가 문장 분석 종료 조건(Stop-condition)으로 사용할 수 있는 또 하나의 기준이 된다. 검사 단어를 기준으로 문장 분석을 하던 중 '용언+연결형 어미'가 나오면 문장 분석을 종료할 수 있다. 그러나 일부 용언은 동일 주어나 동일 목적어를 가진 문장의 서술어가 될 수 있다. 이것은 용언이 연결형 어미일 때 주로 발생하는데 본 논문에서는 연어 오류 단어가 동일 주어일 때와 동일 목적어일 때 문장 분석이 실패하지 않도록 문장 분석 종료 조건에 제약을 둔다. 즉, 필수 문장 요소인 목적어나 주어 다음에 용언+연결형 어미가 올 때 문장 분석을 종료한다.

다음은 연어 오류 단어가 동일 주어인 문장의 예다.

[예 2-1] 그는 얼굴이 일그러지고 매우 불거졌다(X).
(연어오류단어) (검사단어)
=> 그는 얼굴이 일그러지고 매우 붉어졌다(O).

'그는 얼굴이 일그러지다'와 '그는 얼굴이 매우 붉어졌다.'의 두 문장이 연결형 어미 '-고'로 대등하게 연결되어 '얼굴'이라는 같은 주어를 사용하는 문장이다. 검사단어 '불겨졌다'부터 연어 오류 관계인 주어를 찾는 과정에서 문장 분석 조건이 될 수 있는 '일그러지고'가 나오면 문법 검사기는 필수 문장 요소인 목적어 또는 주어가 이미 나왔는지 확인한다. 검사단어에서 왼쪽 방향으로 검색을 시작하여 아직 '불거지다'의 주어가 될 수 있는 단어를 발견하지 못하였으므로 계속 문장 분석을 하여 '일그러지다'와 '불거지다'의 공동 주어인 '얼굴이'를 찾는다. '얼굴'이라는 명사는 '불거졌다'의 용언과 의미상 연어 오류 관계에 있으므로 '불거졌다'를 '붉어졌다'로 교정한다. 동일 목적어를 가지는 유형도 같은 방법으로 교정한다.

다음은 연어 오류 단어가 동일 목적어인 문장의 예다.

[예 2-2] 그는 진정으로 잘못을 뉘우치고 회계했다(X).
(연어오류단어) (검사단어)
=> 그는 진정으로 잘못을 뉘우치고 회개했다(O).

용언 '회계했다'와 목적어인 '잘못을'은 서로 연어 오류 관계이다. 검사단어 '회계했다'는 돈의 나가고 들어오는 것에 대한 셈이나 금전의 출납에 관한 사무를 보는 것을 의미한다. 검사단어 '회계했다'를 기준으로 앞 방향으로 분석을 하던 중 '뉘우치고'가 나오면 동일 목적어를 가질 수 있는 활용 형태인지 검사한다. 여기서 필수 문장 요소인 목적어나 서술어가 아직 발견되지 않았고 동일 목적어를 가질 수 있는 서술어이므로 계속 문장 분석을 진행한다. 다음 단어 '잘못을'이 '회계했다'의 연어 오류 관계를 가지므로 '회계했다'를 '회개했다'로 교정한다.

다음의 예는 위의 예와 다르게 서술어가 동일 주어나 동일 목적어를 가지지 않는 예이다.

[예 2-3] 형사가 증거물을 제시했기에 (범인은) 죄를 인정했다.

[예 2-4] 아들은 가방을 가지러 학교에 갔다.

[예 2-3]은 어미 '-기에'로 연결된 문장이다. '죄를 인정했다'의 주어와 '증거물을 제시했다'의 주어는 서로 다르다. 일반적으로 어미 '-기에'는 선행하는 문장의 주어와 뒤에 오는 문장의 주어가 다르다. [예 2-4]는 어미 [-러]로 연결된 문장이다. 어미 '-러'는 선행하는 문장과 뒤이어 오는 문장이 보통 같은 주어를 가진다. 위의 예에서 보듯 의미상 동일 주어를 가질 수 있는 문장을 해결하는 방법으로 접속어미의 통사적 성질을 이용할 수 있다[13]. 어떤 어미는 선행하는 절과 뒤이어 오는 절이 공통된 주어를 요구하고 어떤 접속 어미는 그렇지 않다.

동일한 주어나 목적어를 가진 문장은 위에서 설명한 부분 문장 분석으로 의미 오류나 문체 오류를 처리할 수 있다. 그러나 다음과 같은 문장은 부분 문장 분석 기법으로 처리하기가 어렵다.

[예 2-5] 수업이 바뀐다고 해서 학생들에게 좋은 영향을 미칠지는 모르나, 수업이 작은 출발점이 될 수 있다.

=> 수업이 적은 출발점이 될 수 있다(X).

[예 2-6] 수업이 작은 반이 성적이 낮다(X).

=> 수업이 적은 반이 성적이 낮다(O).

[예 2-5]는 '작은'이 '출발점'의 서술어인 반면 [예 2-6]에서는 '작은'은 '수업'과 '반'을 동시에 서술하고 있다. 그러나 이런 '수업이 작다'는 단일 문장으로 보면 연어 오류 관계에 있다. [예 2-5]는 '작다'가 '수업'의 서술어 역할을 하지 않으므로 연어 오류 관계가 없으나 [예 2-6]은 아니다. 이 같은 복합 성분을 띤 것은 부분 문장 분석에서 예외로 처리한다.

위의 예에서 본 것 외에 부분 문장 분석이 다수 어절 내에서 분석에 실패하는 가장 큰 원인은 부분 문장 분석 과정에서 어휘 정보가 여러 개 존재하는 중의적인 단어를 처리할 때다. 지배소와 피지배소의 의존관계를 결정할 때, 문장 분석을 종료하려고 현재 어절이 '용언+종결형'인지 결정할 때 어휘 정보가 중의적이면 그 결과를 예측하기 어렵다. 본 논문에서는 이 문제를 해결하기 위해 어휘 중의성을 유형별로 분류하여 문법 검사기의 검사 효율을 높일 수 있는 방법을 제안한다.

2.2 어휘 중의성 유형

부분 문장 분석에 실패하는 주된 원인은 다수 어절 내에 존재하는 어휘적 중의성 때문이다. 이 어휘적 정의성을 해결하기 위해 자연언어를 연구하는 여러 분야에서 많은 노력이 있었다. 어휘 중의성에는 동형 이품사와 이형 동품사가 있다[12]. 문법 검사기의 부분 문장 분석 과정에서 의존 관계를 결정하기 위해 각 어절의 형태소 정보와 의존 규칙을 바탕으로 입력 문장의 각 어절에 대한 의존 관계를 설정한다. 이 때 동형 이품사 때문에 구조적 중의성이 많이 발행한다. 본 논문에서는 의미·문체 검사기가 부분 문장 분석을 수행할 때 문제가 되는 어휘 중의성을 중심으로 중의성 제거 방법을 연구하고, 축적된 결과로 중의성 제거 규칙을 확장한다. 다음은 문법 검사 과정에서 가장 자주 나타나는 어휘 중의성의 예를 세 가지로 분류하여 아래에 소개한다.

2.2.1 동사와 명사

동사의 어간과 어미가 결합한 형태가 명사와 형태상으로 동일한 단어를 말한다.

문법 검사기가 '가격이 사다'는 문장을 검사하면, '가격이'와 '사다'는 주어와 용언 사이에 의미상 연어 오류 관계가 있다고 판단한다. 따라서 '가격이 싸다'로 교정한다. 그러나 '사다'가 용언의 활용된 형태 중에서 '사실'은 명사 정보 역시 가지고 있다. 그러므로 [예 2-8]에서는 '사실'을 '싸실'로 교정하지 않는다.

동사와 명사) 사실

-> 동사 : '사'(어간) + 시(존칭 선어말 어미) + ㄹ(관형형 어미)
-> 명사 : '사실' (실제 존재하는 일)
[예 2-7] 자동차 가격이 사다
[예 2-8] 저문은 그가 범인이라는 명백한 사실

2.2.2 용언과 부사

'졸라'의 어휘 중의성 해결은 경험적 규칙에 따른다. '졸라'가 용언으로 쓰일 때는 '조르다'의 불규칙 변형으로 '동이거나 감은 것을 단단히 죄다'와 '차지고 끈덕지게 요구하다'의 뜻을 가지고 있는 용언으로 사용하는 단어이다. 그러나 인터넷의 보급으로 생긴 새로운 사회 문제 중 하나로써 극단적인 언어파괴의 예인 '매우'와 '아주'의 뜻을 가진 속어로 사용될 때가 있다.

용언과 부사) 졸라

-> 동사 : '조르'('르'불규칙) + (어미 '아/어' 동반)
-> 부사 : '아주', '매우'의 뜻을 가진 속어
[예 2-9] 아버지를 졸라 장난감을 샀다.
[예 2-10] 야! 그 영화 졸라 재밌다.

[예 2-10]은 언어파괴의 예이다. 속어이기 때문에 사전에는 나와 있지 않으나, 통신이나 대중 매체 상에서 흔히 접할 수 있는 잘못된 단어이다. [예 2-9]는 '조르다'의 연결형으로 쓰인 예이다.

2.2.3 동사와 조동사

동사와 조동사) 하다

-> 보조용언 : (어미 '-여야/아야'를 동반)
-> 본용언
[예 2-11] 이번의 정책은 새로운 방법으로 보아야 한다.

위의 예처럼, '하다'는 본동사로도 쓰이지만 앞에 '-어야/아야'와 같은 어미가 오면 보조 용언의 기능을 한다. 이 때, 보조용언 '-어야/아야 하다'라는 패턴 정보를 규칙에 넣지 않으면, '하다'의 피지배소는 '정책'으로 간주한다. 그런데 '정책을 하다'는 '정책을 펴다'의 연어 오류 규칙으로 분석되어 '정책을 펴다'로 잘못 교정된다. 이와 같아 '-어야/아야 한다', -ㄹ/을 수 있다'와 같이 동시에 본용언과 보조 용언으로 쓰일 수 있는 동사는 표현 자체를 하나의 관용어구로 묶어서 처리한다.

2.2.4 명사와 조사

명사와 조사) 밖

-> 명사 : 밖에 있다.
-> 조사 : 밖에 없다.
[예 2-12] 이사짐은 집 밖에 내어 놓아라.
[예 2-13] 최선의 방법은 공부밖에 없다.

'밖'은 앞의 명사와 결합할 때 의미상 품사가 달라진다. [예 2-12]는 '집 바깥에'의 의미로 해석되어 명사가 된다. [예 2-13]은 '없다'와 어울려 '최선의 방법'을 한정한다. 따라서 '밖'은 조사로써 '공부'와 결합하여 '공부밖에'가 된다.

2.3 문법 오류를 처리할 때 어휘적 중의성에 의해 발생하는 문제점

어휘 중의성이 있는 단어를 처리할 때 발생하는 문제는 문법 검사기가 현재 어느 위치에 있는가에 따라 다르다. 현재 검사하고 있는 검사 단어의 어휘가 중의적일 때 발생하는 '규칙 전처리 오류'와 문장 분석 과정에서 분석 중인 단어의 어휘가 중의적일 때 올바른 문장을 비문으로 간주하거나 단문을 이중 문장으로 간주하여 연어 오류 단어를 찾기 전에 문장 분석을 종료하는 '문장 분석 오류'가 있다.
다음은 '문장 분석 오류'의 규칙이다.

문장 분석 오류 규칙
규칙 1 (품사 혼동 오류)
검사 단어 : 밖
연어 오류 단어 : 있다/없다
대치 규칙 : 앞 명사와 붙이기
[예 2-14] 자동차는 길밖에 있다(X).
=> 자동차는 길 밖에 있다(O).
[예 2-15] 시외로 가는 그 길 밖에 없다(X).
=> 시외로 가는 그 길밖에 없다(O).

[예 2-14]와 [예 2-15]은 품사의 혼동으로 오류가 있는 문장이다. '밖에'가 뒤의 용언 '있다'와 결합하면 '바깥에'라는 '명사+조사'로 분석된다. 따라서 명사(길)와 명사(바깥)으로 형태소 분석되어 띄어써야 한다. [예 2-15]은 '밖에'가 용언 '없다'와 의미상 결합하여 앞의 명사 '길'을 한정하는 조사가 되어 앞의 명사와 붙여써야 한다. [예 2-14]와 [예 2-15]은 뒤의 용언에 따라 '밖에'의 품사가 변해 어휘적 중의성을 가져오는 예이다.

문법 검사기는 맞춤법 검사와 의미·문체 검사를 순서적으로 수행한다. 맞춤법 검사 과정에서 현재 어절이 의미·문체 검사를 할 수 있도록 어절의 어간과 형태소 분석 정보를 저장한다. 어휘 중의성이 있는 검사 단어는 여러 개의 표제어와 어휘 정보를 가질 수 있다. 맞춤법 검사기는 여러 개의 표제어 중 가장 짧은 표제어와 이 표제어를 어간으로 한 형태소 정보를 의미·문체 검사기로 넘긴다. 의미 문체 검사기의 규칙 검색부는 어절의 어간(Root)을 표제어로 규칙베이스에서 규칙을 찾는다. 그러므로 어간의 길이가 다르게 분석되는 어절은 규칙 처리가 힘든다. 이를 해결하려면 서로 다르게 분석되는 모든 어간에 대해 규칙을 적용해야 한다. 규칙에는 의미나 문법 오류 유형에 따라 검사단어의 품사가 명시되어 있다. 문법 검사기의 규칙 검사부는 규칙에 명시한 검사 단어의 품사와 현재 검사 단어의 품사가 같은지 검증한다. 이 때 현재 검사 단어의 어휘 중의성을 제거하지 않으면 규칙 검사부의 검사 효율이 떨어진다. 이를 '규칙 전처리 오류'로 구분한다. 다음은 규칙 전처리 오류의 예다.

규칙 전처리 오류
규칙
검사 단어 : 하다
검사 단어 형태소 정보 : 동사
연어 오류 단어 : 정책
대치 단어 : 펴다
[예 2-16] 새로운 장관은 수출 지향 정책을 했다.

[예 2-16]는 검사단어 '하다'가 본용언으로 사용되었으므로 '정책'과 '하다' 사이에 의미 오류가 발생했다. 앞의 [예 2-11]은 '하다'가 보조용언으로 사용되어 '정책'의 직접 서술어가 아니므로 '정책'과 '하다' 사이에 의미 오류가 없는 문장이다. 이 두 문장을 처리할 때 '하다'의 어휘 중의성을 해소하지 않으면 문법 검사기가 [예 2-11]을 '이번의 정책은 새로운 방법으로 보아야 편다'로 잘못 교정하게 된다.

3. 문법 검사기에서 확장된 지식 베이스를 이용한 중의성 해결 방법

현재 개발된 대부분의 맞춤법 검사기에는 품사 태깅 시스템이 부착되어 있지 않다. 부산대학교에서 시험적으로 품사 태깅 시스템을 구현하였으나[22], 본 논문에서는 품사 태깅 시스템 없이 의미 오류나 문체 오류를 검사하는 과정에서 정리된 경험적 규칙을 이용하여 어휘 중의성을 제거하고자 한다. 본 논문에서 사용하는 확장된 경험적 규칙은 좌·우 어절의 어휘 정보를 이용하여 현재 어절의 형태소 정보를 구한다. 이는 한국어에서 자주 나타나는 어절 간 수식 관계를 분석하여 구한 정보다.

문법 검사기에서 어휘 중의성을 해결하려는 목적은 문법 검사 처리 성능의 향상이다. 한국어에서 어휘 중의성을 제거하는 방법은 확률을 이용하는 방법, 퍼지망을 이용하는 방법, 신경망을 이용하는 방법 등 다양하다. 이 같은 중의성 제거 방법을 사용할 때 어휘 중의성 제거율은 높아지지만 문법 검사기에 적용하는 구현이 어렵고 문법 검사기의 검사 속도를 떨어뜨릴 수 있다.

3.1 중의성 제거 규칙 사용

문장 분석 과정에서 형태소적 중의성이 있는 단어를 만나면 어절 간 의존 관계를 결정할 때 예측하기 어려운 결과가 나올 수 있다. 본 논문에서는 문장 분석 과정에서 중의성 제거 규칙을 이용해 중의성을 제거한다. 한국어에서 인접한 두 어절을 살펴볼 때, 앞 어절의 어말 부분이 뒤 어절과 문맥적 관련성이 깊고 뒤 어절의 어두 부분이 앞 어절과 문맥적 관련성이 깊다. 한국어의 이런 특성 때문에 어말-어두 공기 정보를 이용한 문맥 확률을 이용하여 어휘 중의성을 해소할 수 도 잇다. 본 논문에서는 문법 검사기나 의미나 문체 오류를 처리할 때 자주 문제가 되는 어휘 중의성 유형을 중심으로, 이를 해결할 수 있는 문맥 정보를 경험적 언어 지식에 의해 만들었다. 이를 "중의성 제거 규칙"이라 한다.

본 논문에서 사용하고 있는 규칙의 유형은 어휘별 중의성 제거 규칙과 일반적 중의성 제거 규칙으로 분류할 수 잇다. 어휘적 중의성 제거 규칙은 입력 문장의 각각

의 어휘에 대해 규칙이 존재한다. 예를 들면 앞 어절의 어미가 '-어야', '-야야', '-해야'로 끝난 용언일 때 '하다'는 보조용언이다. 그리고 일반적 중의성 제거 규칙은 모든 입력 문장에 대해 동일하게 적용되는 규칙으로 먼저 분석된 왼쪽 또는 오른쪽 어절의 품사 정보를 이용해서 현재 어절의 품사 정보를 결정한다. 예를 들면 왼쪽 문장 분석을 할 때 먼저 분석된 어절이 용언이고 왼쪽에 이웃한 어절이 동사+관형형, 명사+목적어로 중의적 어휘 정보를 가질 때 '동사+관형형'보다는 '명사+목적어'를 우선 선택한다. 규칙을 적용하는 과정은 중의성이 있는 어절을 중심으로 좌우 문맥 정보를 보고 중의성을 제거한다.

한국어에서 어두는 매우 다양하므로 너무 세분화하면 규칙의 유형이 실제 문장에서 나타나지 않을 수 있다. 따라서 본 논문에서는 중의성 제거 규칙을 설정할 때 사용하는 어두의 범주를 명사, 본용언, 보조용언, 술어부사, 부사, 관형사, 수사 7가지로 제한한다. 그러나 어말은 그렇게 다양하지 않으므로 범주를 구체적으로 세분화해서 사용해도 규칙의 유형이 문장에 나타나지 않을 가능성이 작다.

가) 어휘별 중의성 제거 규칙(Specific Disambiguation Rule)

어휘별 중의성 제거 규칙이란 문장에서 나타나는 각각의 어휘에 대해 중의성 제거 방법이 존재하는 규칙이다. 어휘별 중의성 제거 규칙은 다음과 같은 형태로 구성된다.

RULE(#)=(WORD, Direct, Condition, Selection)

어휘별 중의성 제거 규칙은 규칙에 명시된 특성 어휘가(WORD) 나타날 때만 적용된다. 중의성이 있는 특정 어휘(WORD)가 입력되면 이웃한 왼쪽 어절이나 오른쪽 어절이 규칙에서 제시한 조건(Condition)을 만족하는지 검증한다. 조건을 만족하면 현재 단어는 Selection에 명시한 품사로 해석한다.

예로 용언 '하다'가 본용언과 보조 용언으로 해석될 때 중의성을 제거하는 규칙과 '되도록'이 동사+어미와 부사로 해석될 때 중의성을 제거하는 규칙은 다음과 같다.

```
RULE(1)=(하다, LEFT, e 어야|e 아야|e 야, PX)
RULE(2)=(되도록, LEFT, j 이 | j 가, PV)
RULE(3)=(되도록, RIGHT, MM|MA, MA)
  LEFT : 왼쪽 어절
  RIGHT : 오른쪽 어절
 (E:어미, J:조사, MM:관형사, MA:부사, PX:보조용언, PV:본용언)
```

본용언 '하다'는 정책과 연어 오류 관계이다. '정책을 하다'보다는 '정책을 펴다'로 쓰는 것이 바람직하다. 그러나 [예 2-11]에서 사용된 '하다'는 왼쪽 어절의 어미가 '-아야'가 나왔기 때문에 보조 용언으로 해석되어 '정책'과 연어 오류 단어가 될 수 없다. 그러나 [예 2-14]는 '하다'가 본용언이고 연어 오류 단어 '정책을'이 발견되어 '하는'을 '펴는'으로 교정한다.

나) 일반적 중의성 제거 규칙(General Disambiguation Rule)

이 규칙은 문장 분석을 할 때 이미 분석된 어절의 품사를 이용해서 그 다음 입력된 어절의 품사를 결정하는 방법이다. 즉, 검사 단어를 시작으로 문장 분석 방향에 따라 입력된 어절의 어휘가 중의적일 때 앞에 분석된 어절의 품사를 이용해서 중의성을 제거하는 방법이다. 문법 검사기는 검사단어의 연어 오류를 찾는 과정에서 앞 또는 뒤 방향으로 연속하는 어절의 형태소 정보를 저장하고 있다. 즉, 이미 분석된 정보를 이용해 현재 단어의 중의성을 제거하기 때문에 문장 분석 과정에서 중의성 제거에 필요한 시간이 많이 늘어나진 않는다.

일반적으로 중의성 제거 규칙은 다음과 같은 형태로 구성된다.

RULE(#)=(중의성유형, Left, Condition, Col/Anti),(Right, Condition, Col/Anti), Selection)

'중의성 유형'은 중의적으로 분석된 형태소 정보를 나타내고 condition은 중의성을 제거하는데 필요한 왼쪽(Left) 또는 오른쪽(Right) 어절의 품사 정보를 나타낸다. Col/Anti는 주어진 조건과의 관계를 나타낸다. Anti_collocation(Anti)은 왼쪽이나 오른쪽 어절이 Condition에 명시된 품사가 아닐 때 조건을 만족하고, Collocation(Col)은 왼쪽이나 오른쪽 어절이 Condition에 명시된 품사로 분석될 때 조건을 만족한다. 조건이 만족하면 규칙에 의해 Selection에 명시된 품사를 선택한다.

예로 명사+조사와 용언+관형형어미로 해석되는 어절에서 중의성을 제거하는 규칙은 다음과 같다.

RULE(4)

(N+J & P+etm,(Left,0,0),(Right, N, Anti), (N+J)|N);

RULE(5)

(N+J & P+etm,(Left,(P+etm)|N|NN+jcm),Anti),(Right, N|N+j,Col),P+etm);

LEFT : 왼쪽 어절, RIGHT : 오른쪽 어절

N+J : 명사+조사, N+jcm : 명사+소유격조사

N : 조사가 생략된 명사

P+etm : 용언+관형사형어미

현재 어절 오른쪽에 있는 어절이 명사가 아니면 현재 어절을 명사+조사나 조사가 생략된 명사로 본다. 오른쪽 어절이 명사고, 왼쪽 어절이 명사+소유격조사나 명사 또는 용언+관형형어미가 아닐 때 현재 어절을 용언+관형형어미로 본다.

3.2 어휘 중의성 제거 규칙을 이용한 부분 문장 분석

문법 검사기의 규칙 검증 처리부는 의미·문체 오류를 검사하려고 검사 단어를 기준으로 왼쪽, 오른쪽에 연어 오류 관계가 있는 단어를 조사한다. 검사단어와 연어오류 단어 사이에 여러 어절이 있을 때는 문장 분석기에서 의존 문법에 기반한 부분 문장 분석을 한다.

다음은 어휘 분석결과를 이용하여 의존 문법으로 검사 단어의 왼쪽 또는 오른쪽 방향으로 문장 분석을 하는 알고리즘이다.

(3)에서 어휘 분석을 하여 (4)에서는 형태소적인 중의성이 있으면 경험적 규칙을 이용해서 중의성을 제거한다. 처리 단계 (5)~(10)에서 입력 어절의 지배소를 찾아 (11)에서 의존 관계를 분석하여 결과를 저장한다. 새로운 어절의 지배소를 찾을 때는 문장의 투영성[2]을 보장한다. 이미 의존 관계가 설정된 어절은 지배소의 위치를 확인한다. 만약 지배소가 존재하지 않으면 문장 분석을

끝낸다.

```
1) 검사 단어로부터 왼쪽 또는 오른쪽 방향의 I번째 어절(WORD(I))을 가져 온다.
2) if( 의존 관계가 설정 안 된 어절이면 ) {
3) MORP(i)=CHECK-ALL-INFO(I))
4) RMAMBIGIGUITY(MORP(I))
5)    for( IndexOfGov=I-1; IndexOfGov >= 0; ) {
6)       If(MORP(IndexOfGov)와 MORP(I) 의존관계 성립 {
7)             GOV(I)=IndexOfGov
8)               Break;
9)    } else
10)  IndexOfGov=GOV(IndexOfGov)
11)  IndexOfGov=NOT_EXIST
12) } //end for
13) STORE Array(WORD(I), MORP(I), Gov(I))
14) } else
15)      IndexOfGov = RETRIVEArray(I)
16) if(IndexOfGov != NOT_EXIST) return TRUE;
17) else return STOP_NOW
```

[그림 1] 부분 문장 분석 알고리즘

문법 검사기는 하나의 표제어에 대해 여러 개의 규칙을 적용할 수 있다. 규칙에 따라 연관 관계가 있는 어절의 위치가 다르므로 문장에서 분석할 어절 수는 달라질 수 있다. 그러나 이미 의존 관계가 설정된 어절은 다시 분석하지 않는다.

4. 문법 검사기 시스템 구성도

문법 검사기에서 의미 오류와 문체 오류를 검증하는 부분은 규칙 검색부, 규칙 전처리기, 규칙 처리부, 오류 단어 교정부로 구성되어 있다[6]. 규칙 처리부에서는 연어 오류 단어와 검사 단어 사이에 여러 어절이 있을 때 부분 문장 모듈을 호출한다. 어휘 중의성 제거 규칙은 규칙 전처리부와 부분 문장 분석 모듈에서 이용된다.

4.1 의미 오류 처리부의 중심 모듈

(1) 지식 베이스 검색부

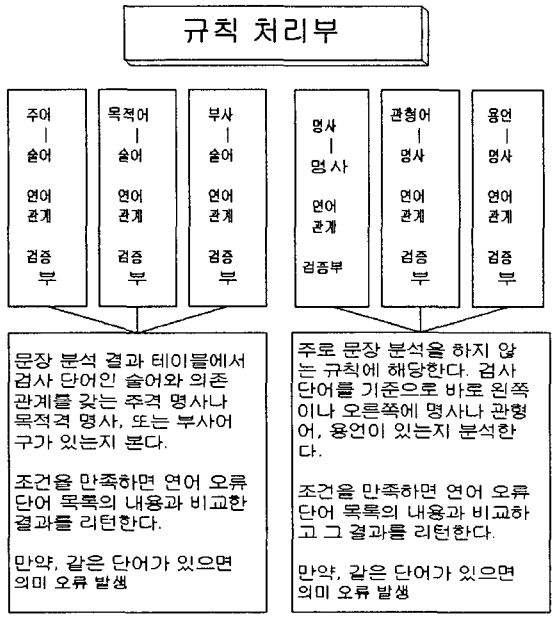
검사 단어에 적용될 N개의 확장된 의미 오류 처리 규칙을 지식베이스에서 찾아 그 시작 위치를 리턴한다.

(2) 규칙 전처리기

규칙을 검사하고 교정하는 과정에서 편리하도록 의미 오류 처리 규칙 내용을 구조체 변수에 대입하난. 규칙에서 제안하는 검사 단어의 어휘 정보와 현재 처리 중인 검사 단어의 어휘 정보가 일치하는지 검증한다. 이 때 검사 단어의 어휘가 중의적이면 중의성 제거 규칙으로 검사 단어의 어휘 중의성을 제거한다.

(3) 규칙 처리부

규칙 전처리기에서 세팅한 구조체 값에 따라 연어 관계가 있는지 검색하고 있으면 그 단어가 연어 오류 관계 단어인지 본다. 규칙 처리부는 연어 오류 단어의 문장 성분에 따라 크게 6개의 검증부로 구분된다 [그림 2]는 규칙 처리부의 세부 모듈을 나타낸다.



[그림 2] 의미 · 문체 규칙 처리부

(4) 오류 단어 교정부

의미 오류가 발생하면 오류 어절(검사 단어)을 대치 단어로 교정한다. 대치 단어와 교정 규칙 정보는 규칙 전처리기에서 설정한 값을 사용한다.

4.2 구분 분석을 위한 모듈 상호 관계

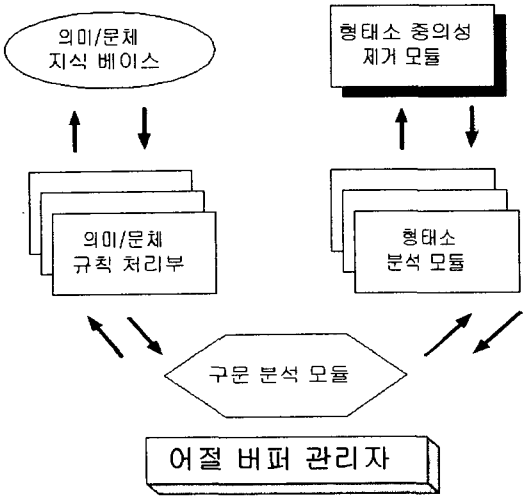
검사 단어와 연어 관계 단어 사이의 문장을 분석할 때 필요한 모듈 및 상호 관계는 [그림 3]에 나타나 있다.

형태소 분석 모듈은 어절의 전자 사전 정보를 중심으로 실질 형태소와 형식 형태소의 품사를 규정한다. 이때 '이다'명사, '하다'명사, '되다'명사 등이 뒤에 '이다', '하다', '되다'와 결합할 때에는 용언으로 분석한다. 마약 N번째 어절의 어휘가 중의적이면 중의성 제거 모듈에서 중의성 제거 규칙을 이용하여 N번째 어절의 어휘 중의성을 제거한다. 형태소 분석 모듈에서 구한 어휘 정보와 의존 문법을 기반으로 지배소와 의존소 간의 의존 관계를 설정한다.

의미 · 문체 규칙 처리부는 부분 문장 분석 결과와 지식베이스에 기반해서 후보 연어 관계 단어(즉, 연어 관계 단어와 같은 품사 정보를 가지는 단어_가 연어 오류 단어인지 비교한다. 의미 오류 또는 문체 오류가 발생하면 지식베이스의 교정 규칙에 따라 교정을 수행한다.

5. 결론

확장된 지식베이스를 이용하여 부분 문장 분석기의 성능을 높인 의미 · 문체 처리의 기본 목적은 가능한 한 문법과 오류가 없으며 이해하기 쉬운 문장으로 교정해 나가는 것이다. 나아가 각 오류에 대한 상세한 도움말과



[그림 3] 부분 문장 분석을 위한 모듈간 상호관계

교정 예를 제시하여 문법 검사기의 사용자가 우리말과 글에 대한 정확한 사용법과 의미를 학습하고 그 중요성도 인식할 수 있도록 한다. 이를 위해서 경험적 규칙을 바탕으로 하여 부문 문장 분석기에서 이용하는 지식베이스를 더욱 확장한다. 그러나 한 어절 단위의 맞춤법 오류나 띄어쓰기 오류를 처리하는 맞춤법 검사·교정기의 기능과는 달리 여러 어절을 이용해서 오류 분석을 하는 문법 검사기는 문장 분석을 해야 한다. 본 논문에서는 말뭉치를 통해 얻어진 경험적 규칙을 확장, 정리하여 지식 베이스의 활용 범위를 넓히고, 확장된 지식베이스를 이용한 의존 문법으로 부분 문장 분석을 이용하여 연어 오류 단어를 찾는다. 문장 분석에서 지배소와 의존소의 의존 관계를 결정하려면 어절의 세부적 형태소 정보가 필요하다. 그러나 한 어절은 여러 개의 형태소 정보를 가질 수 있다. 본 논문에서는 형태소적 중의성이 있는 단어를 이용해 의존 관계를 결정할 때 중의성 제거 규칙을 사용하였다.

의미 오류가 전체 문서에서 나타나는 횟수는 맞춤법 오류나 표준어 오류에 비해 적다. 그러나 의미 오류나 문체 오류는 사용자의 실수로 생긴 오류보다 사용자가 몰라서 발생할 가능성이 큰 오류이다. 문서에 나타난 횟수가 적어도 의미 오류나 문서 오류를 검사 교정 하는 일은 매우 중요한 작업이다.

향후 과제로는 지식 베이스를 좀더 세분하기 위해 명사의 확장된 하위 범주화, 의미적 명사 분류, 한 검사 단어의 두 가지 규칙에서 동시에 연어 오류 단어를 찾았을 때 발생하는 규칙 충돌 문제와 부분 문장 분석으로 처리하기 어려운 복합문에서 연어 오류 단어와 검사 종결 문제 등이 해결되어야 한다. 그리고 앞, 뒤 어절의 품사를 이용한 중의성 제거 규칙은 앞, 뒤 어절의 중의성이 제거되지 않았을 때는 적용할 수 없다는 단점이 있으므로 이 문제가 해결되어야 한다.

참 고 문 헌

[1] 박용욱, 조혁규, 권혁철, 의존 문법을 이용한 한국어 분석기의 구현, 90 정보과학회 봄 학술발표논문집, pp. 191-194, 1990
[2] 손광주, 홍영국, 이종혁, 이근배, 어절간 의존관계 해석을 위한 한국어 파서, HCI 94 발표논문집, pp. 135-136
[2] 홍영국, 이종혁, 한국어 의존 해석을 위한 형태소-통사적 품사 분류 체계, 정보과학회논문지(B) 제 22권 제 9호, pp. 1375-1383
[4] 김영진, 최성필, 손훈석, 박용욱, 권혁철, 단어의 하위 범주화 정보를 이용한 한국어 문법 검사기, 97 인공지능 연구회 춘계 학술발표논문집, pp. 72-75
[6] 김현진, 어절 간 연관 관계를 이용한 한국어 문법 검사기, 정보과학회논문지
[7] 권혁철, 윤애선, 최준영, 단일화 기반 의존 문법에 의한 자연언어 분석 기법, 한국정보과학회 봄 학술발표논문집, 1991
[8] 채영숙, 연어 규칙에 기반한 한국어 문서 교정 시스템의 구현, 부산대학교 박사학위논문, 1998
[9] 김민정, 규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거, 부산대학교 박사학위논문, 1997
[10] 윤준태, 한국어의 대등접속구문 분석, 한국정보과학회논문지, 제 24권 제 3호, pp. 326-335
[11] 이은경, 국어의 접속 어미 연구, 국어연구 제 97호, 국어연구회, 1990
[12] 이상주, 은닉 마르코프 모델을 이용한 두 단계 한국어 품사 태깅, 1994
[13] 강승식, 장병탁, 음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기, 정보과학회논문지, 제 23권 제 5호, 1996
[14] 강재우, 접속 정보를 이용한 한글 철자 및 띄어쓰기 검사기의 설계 및 구현, 한국과학기술원 석사학위논문, 1990
[15] 조영환, 한글 맞춤법 교정기의 설계 및 구현, 한국과학기술원 석사학위논문, 1991
[16] 김덕봉, 최기선, 강재우, 한국어 형태소와 사전-접속 정보를 이용한 한글 철자 및 띄어쓰기 검사기, 언어 연구, 제 26권 제 1호, pp. 87-113, 1990
[17] 채영숙, 김재원, 김민정, 권혁철, 한국어 철자 검색을 위한 형태소 분석 기법, 91 우리말 정보화 잔치 국어 정보과학회, pp. 179-186, 1991
[18] 심철민, 권혁철, 연어 정보에 기반한 한국어 철자 검사기와 교정기 구현, 정보과학회논문지, 제 23권 제 7호, 1996
[19] 박영환, 말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현, 연세대학교 석사학위논문, 1992
[20] 이병훈, 윤준태, 송만석, 말뭉치를 기반으로 한 한국어 철자 교정기의 구현, 한글 및 한국어 정보처리 학술발표논문집, 1993
[21] 소길자, 권혁철, 어휘적 중의성 제거 규칙과 부분 문장 분석을 이용한 한국어 문법 검사기, 정보과학회논문지, 2001.04. pp.305-315
[22] 김광영, 문맥에 의한 중의성 제거와 문장 분석을 이용한 한국어 문법 검사기, 부산대학교 석사학위논문, 2001