부분 문장 분석을 이용한 한국어 문법 검사기 구현*

김현진, 심철민, 권혁철
부산대학교 전자계산학과

Implementation of a Korean Grammar Checker using Partial Sentence Analysis

Hyun-Jin Kim, Chul-Min Sim, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요 약

이 논문은 형태소 사이의 문법 관계(Grammar Relation)에 기반하여 형태소 간의 의존 관계를 규정하고, 이를 바탕으로 의미 오류와 문체를 검증하는 문법 검사기를 제시한다. 이 방법으로 다수 어절에 걸친 의미적 오류 뿐만 아니라 번역체 문구와 뜻의 전달을 어렵게 하는 문구 등과 같이 문장을 힘없게 만드는 문체 오류를 검증한다. 또한 이러한 오류를 검증하기 위한 지식베이스의 구현과 의존 문법(Dependency Structure Grammar)을 이용한 부분 문장 분석 알고리즘을 제시한다. 이 논문에서 제시한 문법 검사기는 향후 파싱 등의 문장 분석에 중요한 자료로 이용될 것으로 기대한다.

1. 서론

기존 철자 검사/교정기는 일반적으로 한 어절이나 좌우의 한 어절을 대상으로 오류를 처리하고 있다[1,2,3,4,5]. 그러나 실제 문서 내에는 철자법으로는 옳으나 좌우 여러 어절에 걸쳐 검증해 보면 찾을 수 있는 의미적 오류나 문법적 오류가 상당수 존재한다[6,7].

아래 [예문-1]과 [예문-2]는 각각 고등학교와 중학교 교과서에서 발췌한 문장으로 '돋구다'가 '돋우다'와 의미적으로 비슷해서 잘못 쓰이고 있는 것을 보여준다.

> [예문-1] 소설 속에 시적인 분위기를 돋구어 주기도 한다. (고등학교 문학 - 문학의 갈래와 그 체계 中)
> [예문-2] 이것은 더위를 타지 않게 하고 입맛을 돋구는 효과가 있다. (중학교 국어 1-1 - 단오 中)

본 논문의 문법 검사기가 다루는 의미적 오류와 문법적 오류는 비슷한 형태나 의미 사용으로 발생한 오류, 문맥상 어색한 단어의 사용이나 시제와 존칭의 오류, 번역체 문구의 사용 등이다.

이런 다수 어절에 걸친 검증이 필요한 오류들을 정확히 교정하려면 통사 분석이나 의미 분석을 행해야 한다[8,9]. 그러나 통사나 의미 분석은 처리 시간과 기억 공간이 많이 소모되므로, 한국어 문서에서 자주 나타나는 오류 유형을 분석해서 이들 사이의 의존 관계와 연관 관계를 이용하면 부분적인 문장 분석만으로도 이런 오류들을 처리할 수 있다.

이 논문에서는 각 형태소 사이의 의존 관계를 정리하고, 부분적인 문장 분석을 통한 효율적인 문법 검사기를 제안한다. 또한 의존 관계를 바탕으로 각 오류에 따른 규칙을 지식베이스화하고, 규칙을 적용해서 다수 어절에 걸친 의미적 오류와 문체 오류를 처리하는 알고리즘을 제시한다.

*본 논문은 과학기술처의 STEP2000과제 중 '국어정보처리 기술개발'사업 과제인 지능형 처리기 개발 과제(주관 연구기관:국어공학센터/시스템공학연구소)의 연구비에 의해 연구되었음.

2. 어절 간의 의미적, 문법적 오류

이 논문에서 제시하는 문법 검사기의 분석 대상은 크게 의미 오류와 문법과 문체 오류로 구분할 수 있다. 의미 오류는 주로 발음이나 철자법, 또는 의미를 혼동하기 쉬우므로 범하는 오류이고, 문법과 문체 오류는 철자법이나 의미상으로는 틀렸다고는 볼 수 없으나 우리말의 어법이나 정서에 맞지 않는 표현이다.

의미 오류 유형	예제 문장
적절하지 못한 낱말의 사용	영희는 매우 공부하였다.(X) 영희는 열심히 공부하였다. (O)
발음상의 오류	그는 입사한 이후 일에 메여 있다.(X) 그는 입사한 이후 일에 매여 있다.(O)
철자법의 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	['하다'와 결합할 수 있는 명사, 지역명]+일생(X) : '일생'과 무정명사와의 결합은 무의미하다.

[표-1] 어절 간의 의미 오류 용례

문법과 문체 오류 유형	예제 문장
영어 번역체 문장	이 물건을 살 필요가 있다.(X) 이 물건을 사야 한다. (O)
일본어 번역체 문장	문학에 있어서 자연이란 소재에 불과하다. (X) 문학에서 자연이란 소재에 불과하다.(O)
겹말 표현의 오류	함께 동행을 하다.(X) 함께 가다.(O)
반복된 문구	이번 전쟁에 대한 성공담에 대하여 한 번 들어보자. (X) 이번 전쟁의 성공담에 대하여 한 번 들어 보자. (O)
바람직하지 못한 문장	그의 존재는 뒷전이었다. (X) 그의 존재는 중요하지 않다.(O)
존칭어 오류	질문이 계신 분은 손을 들어 주세요. (X) 질문이 있으신 분은 손을 들어 주세요. (O)
어려운 한자말 사용 문장	이번 실패는 게으름에서 기인하고 있다고 생각한다.(X) 이번 실패는 게으름 때문이라고 생각한다. (O)

[표-2] 문법과 문체 오류 용례

이 논문의 문법 검사기가 처리하는 의미 오류와 문법, 문체 오류를 간단히 분류하면 앞의 [표-1]과 [표-2]로 정리할 수 있다.
사용자가 많이 범하는 오류를 다룬 한국어 문법과 문체에 관한 참고 도서[10,11,12,13,14,15,16,17,18]들을 중심으로 유형 분류를 하였다.

3. 형태소 사이의 의존 관계

앞 장에서 살펴본 오류를 처리하려면 문장의 통사 분석과 의미 분석을 해야 한다. 그러나 통사 분석과 의미 분석을 하면 오류를 정확히 찾을 수 있으나, 속도가 느리고 많은 공간이 필요하므로 효율적이지 못하며, 더구나 기존 기술로 완벽한 통사 분석과 의미 분석은 불가능하다[8,9,18].

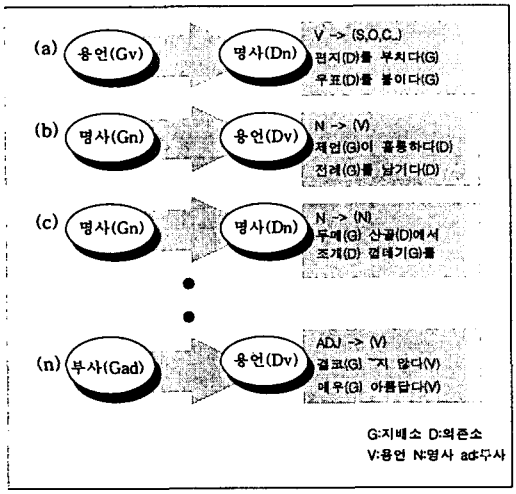
이에 반해 형태소 사이의 의존 관계를 이용하여 오류를 규칙화해서 지식베이스로 구축하면 부분적인 문장 분석만으로도 다수 어절에 걸친 의미적 오류나 문법, 문체 오류를 효율적으로 검증할 수 있다[7].

3.1. 지배-의존 관계의 이용

이 논문에서 제안한 문법 검사기는 기존 통사 분석을 위한 단어의 지배-의존 관계 규칙과는 조금 다른 지배-의존 관계를 이용한다. 문법 검사기에서는 오류 유형에 맞추어 제한된 범위의 문장 분석을 요하므로 오류 문장의 형태에 따라 지배-의존 관계가 효율적으로 변형되어야만 하기 때문이다[19].

아래 [그림-1]에서 보면 (a)에서는 동사가 지배소가 되고 명사가 오류를 인식하게 하는 의존소가 되는 반면, (b)에서는 (a)와 반대의 형태가 된다. 문법 검사기에서 쓰이는 의존 관계는 검증의 효율성을 위해 오류가 발생 가능한 형태소가 지배소가 되고, 이와 연관 관계가 있는 형태소들이 의존소가 된다.

[그림-1]의 예에서도 (a)의 '부치다'가 '붙이다'와 의미적으로 혼동하기 쉬우므로 이를 지배소로 보고, 한 문장에서 '부치다'와 연관 관계(Collocational Relation)가 있는 의존소들을 발견하면 의미 오류가 발생한 것으로 인식한다. 즉, 이러한 지배-의존 관계를 기본 구조로 하고, 오류 검증은 단어 간의 연관 관계(Collocational Relation)를 이용한다. 문법 검사기에서 이용하는 단어 간의 연관 관계의 예는 부록 [표-4]에 나와 있다.



[그림-1]문법 검사기에서의 지배-의존 관계 예

3.2. 의존 관계를 이용한 지식베이스 구축

이 논문에서는 앞에서 설명한 형태소 간의 지배-의존 관계와 단어 간의 연관 관계를 기반으로 지식베이스를 구축하고 있다. 아래 [그림-2]는 문법 검사기의 지식베이스 구조이다. 문법 검사기에서 지배소를 발견하면 지식베이스에 저장되어 있는 조건에 따라 검증을 한다. 지식베이스에는 지배소의 조건과 의존소의 위치 정보, 의존 관계, 연어 관계가 오류 결정 정보로 구축되어 있다. 부록의 [그림-6]은 "아무리 강조해도 지나치지 않다"라는 문장을 영어 번역체 문장으로 인식해서 문체 오류로 검증하는데 필요한 지식베이스를 예로 보여 주고 있다.

```
Rule-Number n
[표제어1]{ [대치어1]{ (도움말 번호)(교정 조건)(현재 단어 조건)
                      (검색 방향)(검색 어절 수)(연어 관계)
                      [오류 검증 조건]{ (격 조건1, 오류 결정 단어1),
                                        (격 조건2, 오류 결정 단어2),
                                                  :
                                        (격 조건N, 오류 결정 단어N),
                      }
          }
                                                  :
          [대치어n]{ (도움말 번호)(교정 조건)(현재 단어 조건)
                      (검색 방향)(검색 어절 수)(연어 관계)
                      [오류 검증 조건]{ (격 조건1, 오류 결정 단어1),
                                        (격 조건2, 오류 결정 단어2),
                                                  :
                                        (격 조건N, 오류 결정 단어N),
                      }
          }
}
```

[그림-2] 지식베이스의 구조와 예

4. 부분 문장 분석을 이용한 오류 처리 알고리즘

4.1. 부분 문장 분석

이 문법 검사기에서는 의존도가 있는 단어쌍을 중심으로 제한된 범주 내에서 문장 분석을 시도한다. 즉, 지배소를 중심으로 규칙이 제시한 대로 의존소를 검증하는 과정에서 제한된 형태의 문장 분석을 한다.

본 연구에서는 어절 사이의 오류를 검증하기 위해서 어절에 부여되는 범주를 크게 지배소에서 좌측으로 검증하는 좌범주, 우측으로 검증하는 우범주로 나누고, 지배소의 문법적 상태와 의존소의 관계에 따라 검증하는 어절 수는 제한하였다.

이 논문의 문법 검사기가 제시하는 문장의 범주와 문장 성분에 따른 제약 조건은 다음 [표-3]과 같다. 여기에 있는 제약 조건은 대부분 일반적으로 문장에서 자주 나타나는 경우이며, 그 외의 특수한 경우에는 지배소와 의존소의 오류 규칙과 각 오류 정보에 따라 그 범주와 위치 정보가 다르다.

지배소	의존소	범주 및 최대 위치 정보	무시하는 정보
V+PE	O,S,C	Left - 3	ADJ, AD
V+GE	O,S,C	Right - 2	G
N+J	V+GE	Left - 2	G
N+J	V+PE	Right - 3	ADJ, AD
N+J	N+GJ	Left - 1	none
N+J	N+J	Left - 1 Right - 1	none
V+E	ADJ	Left - 2	ADJ
V+E, N+J	E,J	오류 규칙과 정보에 따라 다름	오류 규칙과 정보에 따라 다름

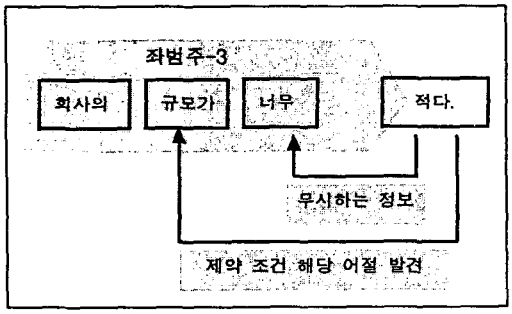
V:동사,형용사 PE:평서형 어미 GE:관형형 어미 E:어미
O:목적어 S:주어 C:보어 J:조사 N:명사 ADJ:부사
AD:부사어 GJ:관형격조사 Left: 좌범주 Right:우범주

[표-3] 문장의 범주와 위치 정보에 따른 분류

[표-3]은 이들 문형 사이에서 지배소와 의존소가 서로 떨어져 있을 것으로 추정되는 범주를 좌범주, 우범주로 나누고, 문법 검사기가 검증할 최대 어절 수를 도표로 나타낸 것이다. 표에서 '무시 하는 정보'는 주로 문장의 부속성분으로서 지배소와 의존소 사이에 존재할 수 있는 어절을 말한다.

[표-3]에 나타난 문장의 범주와 위치 정보를 바탕으로 부분 문장 분석을 하는 예를 [그림-3]에 제시하고 있다.

[그림-3]에서 '적다'는 지배소로서 평서형이므로 의존소인 목적어 '규모', '금액', '키' 등이 문장에 있는지를 조사한다. 이 때 오류 규칙에 따라 지배소인 '적다'에서부터 좌범주로 최대 3어절이 설정되며, 가장 근접한 어절에서 좌측으로 검증을 시작한다. '너무'는 문장 성분이 부속 성분인 부사이므로 검증을 계속할 수 있다. 다음 어절인 '규모가'를 검증하면 이 문장에 의미적 오류가 발생함을 확인할 수 있으며, 따라서 '적다'를 '작다'로 교정하게 한다.



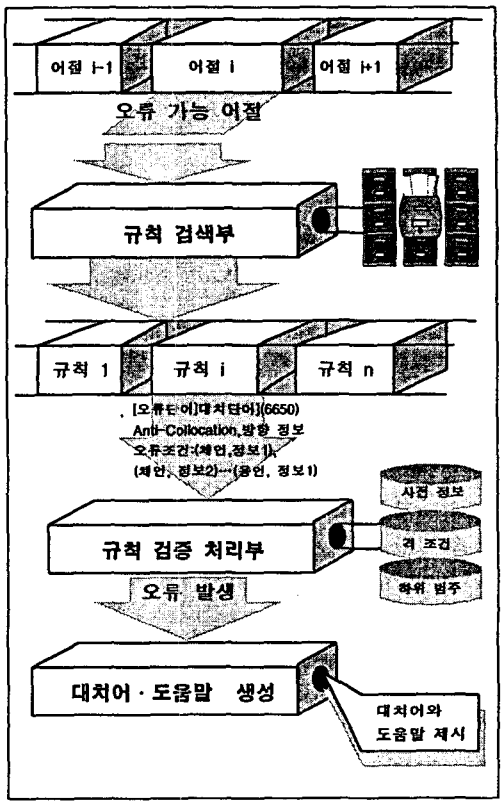
[그림-3] 부분 문장 분석 예

본 논문의 부분 문장 분석은 기본적으로 단문에서 나타나는 문법적 관계만을 고려하고 있다. 그러나 이 연구가 제한된 지배소와 그와 의존 관계가 있는 의존소들을 대상으로 하고 있으므로 본 논문에서 제시한 문장 분석만으로도 오류 검증이 가능하다. 연세대 말뭉치에서 연속된 5어절에서 나타나는 지배소와 오류 규칙에 나타난 의존소와의 위치 정보를 검증한 결과 92% 정도는 본 논문에서 제약한 범주 내에 있음을 알 수 있었다.[1]

1) 연세대 코퍼스에서 임의로 추출한 3만 어절 중 문법 검사기가 다루는 지배소와 의존소의 위치 정보를 실험한 결과이다.

4.1. 의미 오류와 문체 오류 검사 과정

[그림-4]는 문법 검사기가 의미 오류와 문체 오류를 검사하는 과정을 그림으로 나타내고 있다.



[그림-4] 오류 처리 과정

문법 검사기에서 의미 오류와 문체 오류를 검증하는 부분은 크게 1) 규칙 검색부, 2) 규칙 검증 처리부, 3) 대치어-도움말 생성부로 구성되어 있다[7]. 부분적인 문장 분석 기법을 이용해서 오류 어절을 검사하는 과정은 다음과 같다.

1) 어절 버퍼 관리자는 환형큐(circular queue)로 구성되어 있으며, 입력 문서로부터 한 문장을 받아서 문법 검사기에 한 어절씩 제공한다.
2) 의미 오류나 문체 오류가 발생 가능한 어절이 발견되면 규칙 검색부를 구동시킨다.
3) 이미 형태소 간의 의존 관계를 기반으로 구축한 지식베이스에는 검색 어절과 연관 관계가 있는 단어들과 문장 분석 패턴이 있으므로 규칙 검색부에서는 이 규칙들을 버퍼에 담고 규

칙 검증 처리부를 호출한다.
4) 규칙 검증 처리부는 지식베이스에서 획득한 규칙들에 대응하는 규칙 적용 알고리즘에 따라 필요한 형태소 분석 정보와 문장 정보를 수집한다. 이때 하위 범주 사전(동물, 식물, 사람 정보 등)과 기타 사전 정보와 격조건(주격, 목적격 등)루틴을 검증하면서 규칙을 적용한다.
5) 규칙 검증 처리부가 의미 오류나 문체 오류를 인식하면 지식베이스에 저장된 대치어 조건과 도움말 생성 규칙과 함께 대치어-도움말 생성부를 구동한다.
6) 대치어-도움말 생성부는 적절한 교정 규칙에 따라 대치어와 도움말을 생성한다.

4.2. 규칙 적용 알고리즘

[그림-4]의 규칙 검증 처리부에서는 지식베이스로부터 받은 규칙에 따라 좌우 검증을 한다. 그런데 하나의 지배소에 의미·문체 오류가 동시에 발생할 가능성이 있는 문장 패턴이 여러 개 존재할 수 있다. 즉, 지식베이스에서 적용할 규칙이 하나 이상일 수 있다. 따라서 규칙을 적용하는 형태에 따라 크게 1) 단일 규칙 적용 2) 병렬적 규칙 적용 3) 계층적 규칙 적용으로 구분할 수 있다. 각 규칙을 설명하면 다음과 같다.

가. 단일 규칙 적용

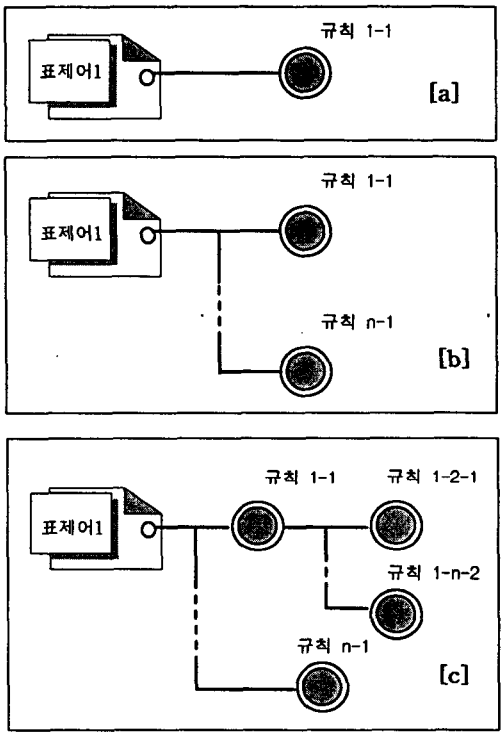
[그림-5]의 (a)에서 보듯이 하나의 지배소에 하나의 규칙이 적용되는 경우이다. 이 경우는 의미·문체가 발생할 가능성이 있는 문장 패턴이 하나인 경우이므로 그 규칙을 그대로 검증해서 오류 발생 여부를 확인하면 된다.

나. 병렬적 규칙 적용

이 적용 방식은 하나의 지배소에 대해 서로 연관이 없는 규칙이 여러 개 존재하는 경우이다. [그림-5]의 (b)에서 보면 한 개의 표제어가 지식베이스에 여러 개의 규칙을 가질 수 있는데 이 규칙들 중에 어느 규칙을 먼저 적용하더라도 검증 결과엔 영향을 주지 않는다. 그래서 규칙들을 병렬적으로 적용해서 먼저 오류가 발생하는 규칙에 따라 대치어와 도움말을 제시하도록 한다.

다. 계층적 규칙 적용

이 경우는 지식베이스의 규칙들 간에 계층이 존재해서, 규칙들 중에 어떤 규칙을 먼저 적용하느냐에 따라 결과가 다르게 나타날 수도 있다. [그림-5]의 (c)에서 보는 바와 같이 하나의 표제어에 각 처리 규칙들을 계층적으로 연결할 수 있다. 비슷한 규칙이 새로 생성되면 다른 규칙들과 다시 계층을 만들어서 새로운 규칙계층을 형성한다.



[그림-5] 규칙 적용

5. 결론

이 논문에서는 통사 분석에서 쓰이는 지배-의존 관계를 연어 관계를 이용한 어절 간 오류 검증에 맞도록 변형했다. 그리고 단어 사이의 연관 관계를 지식 베이스화하였으며 각각의 처리 루틴을 규칙화하였다. 또한 단어 사이의 의미적 오류와 문체 오류의 검증을 최소한의 형태소 분석과 부분적인 문장 분석만으로 가능하도록 했다.

그리고 의미·문체 오류를 검증하는 지식베이스의 구조와 규칙 적용 알고리즘도 살펴보았다. 지식베이스란 지배소와 의존소의 관계를 하나의 규칙으로 만드는 것이라 할 수 있다. 그런

데 지배 관계에 의해 지배소로 결정되는 단어의 개수가 많아지고 이러한 규칙이 계속 추가된다면 지식 베이스의 크기가 엄청나게 커지게 된다. 이 경우는 지배소와 의존소를 바꾸어 지식 베이스를 다시 구축함으로써 지식 베이스의 탐색 키만 바뀌고 처리 결과는 동일하게 하는 결과를 기대할 수 있다[7,19].

우리말 문체 오류는 한 문장에만 그 오류가 국한되지 않고 여러 문장에 걸쳐서 오류가 발견되기도 하므로 문체 오류 지식베이스에 문장에 관한 규칙도 추가하면 이런 오류를 부분적으로 검사할 수 있다.

이 논문에서 제시한 부분적인 문장 분석을 이용한 문법 검사기는 한국어 문체 검사기(Korean Style Checker)로서 의의를 가지며, 형태소 간의 의존 관계는 향후 파싱 등의 문장 분석이나 의미 분석에 중요한 자료로 이용될 수 있을 것으로 기대한다.

참고 문헌

[1] 이병훈, 윤준태, 송만석, "말뭉치를 기반으로 한 한국어 철자 교정기의 구현", 한글 및 한국어 정보 처리 학술발표논문집, pp.285-293, 1993.

[2] 정한민, 이근배, 이종혁, "자판 특성을 이용한 Neuro-Fuzzy 한국어 철자 교정기의 구현", 한글 및 한국어 정보 처리 학술발표논문집, pp.317-328, 1993.

[3] 박종만, "철자 검색기에서 틀린 어절의 처리",한글 및 한국어 정보처리, pp.187-195, 1991.

[4] 강재우, "접속 정보를 이용한 한국어 철자 띄어 쓰기 검사기의 설계 및 구현", 한국과학 기술원 전산학과 석사학위 논문, 1990.

[5] 강승식, 이호석, 문유진, 김영택, "한국어 문법 검사/교정 시스템의 설계", '90 춘계 논문집, 17권 1호, 한국정보과학회, 1990.

[6] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기", 부산대학교 전자계산학과 석사학위 논문, 1995.

[7] 심철민, 김현진, 김영진, 권혁철, "연어 정보를 이용한 한국어 철자 검사/교정기의 성능 개선", 한글 및 한국어정보처리, 한국인지과학회, 1995.

[8] 권혁철, 윤애선, 최준영, "단일화 기반 의존 문법에 의한 자연언어 분석 기법", 한국정보과학회 봄 학술발표논문집, 1991.

[9] 홍영국, 이종혁, 이근배, "의존문법에 기반을 둔 한국어 구문 분석기", 한국정보과학회, 봄 학술발표논문집, 1993.

[10] 미승우, 새 맞춤법과 교정의 실제, 어문각, 1988.

[11] 이인섭, 심영자, 우리말 고운말 1,2, 민문고, 1992.

[12] 이오덕, 우리글 바로쓰기 1,2,3, 한길사, 1992.

[13] 이수열, 우리말 우리글 바로 알고 바로 쓰기, 지문사, 1993.

[14] 원영섭, 초중고 국어 교과서에 나타난 띄어쓰기 용례, 세창출판사,1993.

[15] 김봉모, 국어 정서법, 세종 출판사, 1995.

[16] 박갑수, 국어 문체론, 대한교과서, 1994.

[17] 한국언론연구원, "신문기사의 문체", 한국언론연구원, 1990.

[18] 한효석, 이렇게 해야 바로쓴다, 한겨레신문사, 1994.

[19] Chul-Min Sim, Min-Jung Kim, Hyuk-Chul Kwon, "Automatic Revision of Korean Texts by Collocation Words," Proc. of the 1994 ICCPOL, pp.280-284, Taejon, Korea, 1994.

[20] R.L.Kashyap, B.J.Oommen, "Spelling correction using probablistic method," Pattern Recognition Letters, pp.147-154, 1984.

부록

연어 관계 단어 (Collocation-Relation)	의존 단어들	연어 오류 관계 단어 (Anti-Collocation Relation)
잦다	통행, 시비	많다
타개	위기, 어려움, 정국	타계
제목	논문, 문서, 책, 노래, 영화, 수필	재목
벌이다	논쟁, 결투, 싸움질, 전쟁, 일, 싸움, 사건, 가게, 잔치, 상품, 물건	벌리다
닥치다	재앙, 위기, 어려움, 전쟁	다치다
넓다	범위, 넓이	크다
닿다	손길, 연락, 손, 힘, 신체 명사	닫다
작렬하다	폭죽, 폭발물, 폭탄, 수류탄	작열하다
전래	이야기, 동화, 풍습, 풍속, 습관	전례
외곽	단체, 지역, 도시, 학교, 마을, 건물,시, 도, 지역 정보	외각
엎다	살림, 상, 밥상	업다
제재	가하다	재재, 제제
둘러메다	총, 짐짝, 보따리, 세간, 살림살이	둘러매다
빚	지다, 갚다	빛, 빗

[표-4] 단어 연관 관계(Collocational Relation)

```
Rule-Number 18
[지나치다]( [대치어 없음]
        ( (10024)(지나치 + '~지' 어미)
          (좌 방향 - 2 )(Collocation)
              [오류 검증 조건 1] (부사, 아무리 )
          (좌 방향 - 1 )(Collocation)
              [오류 검증 조건 1] (특정 어미, ~어도),
                                 (특정 어미, ~여도),
                                 (특정 어미, ~아도),
                                 (특정 어미, ~해도)
                                       :
                                       :
          (우 방향 - 1 )(Collocation)
              [오류 검증 조건 1] (용언, 않다 )
        )
)
```

[그림-6] 지식베이스 용례