단어 간 지배 관계 및 연관 관계를 이용한 한국어 교열 시스템

○
심 철민, 김 민정, 이 영식, 권 혁철
부산대학교 전자계산학과

A Korean Revision System Using the governal and collocational relation between words

Chul-Min Sim, Min-Jung Kim, Young-Sik Lee, Hyuk-Chul Kwon
Pusan National UNIV, DEPT. of Computer Science

요 약

스펠러와 같은 오류 처리 기법은 한 어절 사이의 처리에 국한되거나, 또는 수사 처리와 같이 일부 제한된 품사 영역에서만 어절을 넘어선 처리가 행해지고 있다. 한편 교열과 같은 어절 단위를 넘어선 오류 처리는 완벽한 통사 분석과 의미 해석을 반드시 필요로 한다고 생각되어져 왔다. 그리고 현재 한국어 처리에서는 완벽한 통사적, 의미적 처리가 불가능하기 때문에 교열 시스템 또는 어절 단위를 넘어선 오류 처리에 대한 연구가 거의 전무한 실정이다.

본 논문은 어절을 넘어선 오류의 유형을 분류하고, 문장 단위로 관련된 단어 사용 오류를 검사하는 기법과 관련 단어 처리를 위한 규칙 데이타 베이스의 구조를 제안한다. 단어 사이에 존재하는 통사적, 의미적 지배 관계와 연관 관계를 어휘선택 제약으로 이용함으로써 완벽한 통사 분석과 의미 분석이 없이도 교열이 가능하게 하였다.

Ⅰ.서론

현재 구현된 한국어 철자 검사기는 한 단어 내에서 발생할 수 있는 맞춤법 및 띄어쓰기 오류를 검사한다[1]. 철자 검사기는 단어가 철자법상으로 옳지만, 단어와 단어 간의 의미 차이에 의한 문장 단위의 의미적 오류를 처리하지는 못한다. 예를 들어 '검붉은 피가 엉켜 있다.'라는 예문은 철자법상 옳은 문장이다. 그러나 '피가 엉키다'는 표현은 잘못되었다. '엉키다'는 '섞임'의 의미이다. '응고(凝固)'라는 의미로는 '엉기다'라는 단어를 사용해야 한다.

부산대학교 철자 검사기는 예외적으로 수사 처리에서 단어 사이의 오류를 검사한다[2]. 수사는 연속되어 나타나므로 연속된 수사 오류의 처리가 가능하지만 일반적인 한국어 문장에서는 어순이 자유로우므로 단어 사이의 오류 처리를 위해 새로운 방법이 필요하다. 단어 사이의 오류의 대부분은 의미적 오류이며, 이를 처리하기 위해서는 우선 문장의 통사 분석이 선행되어야 한다고 생각되어 왔다. 가장 일반적인 통사 분석 방법에는 파싱이 있다. 파싱은 정확한 의미 분석을 가능하게 하지만 속도가 느리고 많은 공간을 필요로 하므로 실효성이 없다. 일반적으로 사용되는 문서는 문장 구조가 간단하고 일반인들이 범하는 단어 사이의 의미적 오류는 의외로 상당히 유사하다. 그러므로 파싱의 대안으로 일반인들의 빈번한 단어 사이 오류에 대해 규칙 베이스를 이용하여 처리할 수 있다.

본 논문에서는 현재의 철자 검사기가 가지는 한계를 극복하기 위한 한국어 교열 시스템을 제안하며, 교열은 출판물의 잘못된 글자나 문구를 바르게 고치는 것으로 제한한다. 한국어 교열 시스템은 일반인들이 자주 범하는 문구상의 오류들 중 철자 검사기가 처리하지 못하고 있는 것들을 처리하는 것을 목적으로 한다. 본 논문에서는 일반인의 단어와 단어 간의 오류를 발음상의 유사성으로 인한 오류, 의미적 유사성으로 인한 오류, 한자어와 순수 한국어의 의미 중복 오류, 존칭 오류, 명사의 의미적 연관성 오류, 관용구 오류, 띄어쓰기 오류, 그리고 기타 오류 등으로 분류한다. 또한 이렇게 분류된 오류를 규칙 베이스화하여 처리한다. 본 한국어 교열 시스템이 사용하는 규칙 베이스는 단어의 지배 관계에 의해 구성되며, 연관 관계 정보를 가진다. 이 논문에서 제시한 한국어 교열 시스템은 완벽한 통사 분석과 의미 분석 없이 최소한의 형태소 분석만으로 일반인들이 자주 범하는 문구상의 오류들을 처리한다.

Ⅱ.단어 사이 오류 유형 분류

1.교열을 위한 지배 관계와 연관 관계

출판에서의 교열이란 틀린 단어의 교정(校正)과 문구의 교정(校訂)을 포함한 의미이다[3]. 교열 시스템은 철자 검사기의 기반에서 그 처리 영역을 확장하기 위한 시스템이므로 교열의 의미를 다음과 같이 한정한다.

> 교열(校閱)이란, 문장을 구성하는 단어 간의 의미적 불일치 오류나, 여러 단어의 통사 분석에 의해서만 처리될 수 있는 문법적 오류의 검사 및 교정을 의미한다.

문장 교열을 위해 본 논문에서는 지배소(Governor)와 의존소(Dependent) 간의 지배 관계를 사용하여 오류의 유형을 분류하고 처리 방법을 제시한다. <표 1>은 한국어의 일반적인 지배 관계 규칙이다[4]. <표 1>의 지배 관계 규칙은 파싱 등 문장의 통사 분석을 위한 일반적인 규칙이다. 그러나 본 논문에서 구현한 교열 시스템은 일반인들의 빈번한 오류 유형에 한정하여 문장에서 필요로 하는 부분만 최소한의 통사 분석을 행하므로 일반적인 지배 관계 규칙을 그대로 적용할 수는 없다. 일반인들의 오류 유형을 분석한 결과 교열 시스템을 위한 지배 관계에서는 지배소가 변형될 필요가 있다. 예로써 높임말의 경우는 높임의 주체가 지배소가 되고 명사나 용언,조사가 의존소가 된다.

<표 2>에서 교열 시스템을 위해 변형된 지배 규칙을 제안한다.

관계	지배소	의존소
수식	명사	관형사, 명사, 어미, 조사
격부여	조사	명사, 대명사, 수사, 조사, 부사
양상부여	어미	동사어간, 형용사어간, 어미
부가	형용사 어간	조사, 명사, 부사
부가	동사 어간	조사, 명사, 부사, 어미
강조	부사	부사

<표 1> 한국어 지배 관계 규칙

관계	지배소	의존소	예
의미주체	용언	명사	피(D)가 엉기다(G) 실(D)이 엉키다(G) 생각(D)이 얽히다(G)
의미일치	부사	용언	결코(G) -않다(D)
의미연관	명사	명사	노루(D) 꼬리(G) 닭(D) 꽁지(G)
의미연관	본용언	보조용언	가(G) 주다(D)
격지배	용언	조사	-에(D) 얽매이다(G)
의미중복	명사	명사, 용언	봉변(G)을 당하다(D)->변을 당하다 역전(G)앞(D)->역앞 고목(G)나무(D)->고목
어미지배	어미	용언	먹을(D)수록(G)
존칭	명사[주체]	명사, 용언, 조사 선어말 어미[존칭]	아버지(G)께서(D) 진지(D)를 드신다(D)

* G : 지배소(Governor), D : 의존소(Dependent)

<표 2> 교열을 위한 변형된 지배 관계 규칙

한국어 교열 시스템에서 지배소와 의존소를 결정하는데 영향을 주는 요소는 규칙 베이스의 크기이다. 한국어 교열 시스템의 규칙 베이스는 지배소 별로 하나의 규칙을 할당한다. 그런데 지배 관계에 의해 지배소로 결정되는 단어의 개수가 많아지고 이러한 규칙이 계속 추가된다면 규칙 베이스의 크기가 엄청나게 커지게 된다. 지배소와 의존소를 바꿀 경우 규칙 베이스의 탐색 키(Search-Key)만 바뀌고 처리 결과는 같으므로 이러한 경우는 예외적으로 지배소와 의존소를 바꾸어 선택한다.

본 논문에서는 이러한 지배 관계를 기본 구조로 하여, 단어 선택 제약 조건으로 단어 간의 연관 관계를 사용한다. 연관 관계는 하나의 단어가 다의어이거나 동형이의어일 때, 그 의미는 문장 내에서 다른 단어와의 의미 관계에 의해서 결정된다는데 기반한다. 연관 관계는 기계번역 등에서 단어의 올바른 역어 선택 방법의 하나로 사용된다[5]. <표 2>와 같이 변형된 지배 관계 내에서 각 지배소에 대한 의존소를 실제 문서에서 빈번히 오용되는 단어들로 제한한다.

(예 1) 검붉은 피가 엉키다.

(예 1)은 일반인뿐만 아니라 언론사에서도 빈번히 범하는 오류이다. 이 예에는 <표 2>의 '의미주체'에 해당하는 지배 관계가 적용된다. 용언 '엉키다'가 지배소가 되고

명사 '피'가 의존소가 된다. 이와 같이 '엉키다'와 함께 오용되는 명사들은 의존소가 되고, 용언 '엉키다'와 그 명사들의 의미주체 관계에 의해 단어 선택이 올바른 지를 판단한다. 이 예에서 의존소는 지배소와 부정적 관계(NR:Negative Relation)를 가지고 있다. 즉, '엉키다'라는 지배소에 대해 '피'라는 의존소가 나타났을 경우 그 문장은 오류 문장으로 판단된다. 그 반대의 경우는 지배소와 의존소가 긍정적 관계(PR:Positive Relation)에 있다고 한다. 기계번역에서는 긍정적인 연관 관계 밖에 고려하지 않지만, 교열 시스템에서는 오류 검사를 위해 부정적인 관계를 이용하고 대치어 생성을 위해 긍정적인 관계를 사용한다. 그러므로 본 논문에서의 연관 관계는 다음과 같이 정의한다.

> 연관 관계(Collocational Relation)란, 문장의 의미적 오류 처리를 위한 지배소와 의존소 간의 긍정적,부정적 의미 관계이다.

2. 오류 유형 분류

본 논문에서는 단어 간의 지배 관계와 연관 관계의 측면에서 오류 유형을 분류하고 규칙 베이스를 구축한다. 일반 문서에서 범할 수 있는 단어 사이의 오류를 의미 유사성으로 인한 오류, 발음 유사성으로 인한 오류, 존칭 오류, 의미 중복 오류, 관용구 오류, 조사 오류, 용언대 용언의 의미적 지배 오류, 띄어쓰기 오류, 그리고 기타 오류 등으로 분류한다. 각 유형의 예문을 지배 관계 및 연관 관계에 입각하여 도표로 표현한다.

o. 의미 유사성으로 인한 오류

의미는 유사하나 사용되는 상황이 다른 단어끼리 오용되는 경우이다.

(예 2) 총소리가 울리자 노루란 놈은 질겁하여 꽁지가 빠지게 달아났다.

지배 관계		의존소		지배소	
문장	총소리가 울리자	노루란	놈은 질겁하여	꽁지가	빠지게 달아났다
연관 관계를 가지는 단어		닭,제비,새 (새,물고기)		꽁지	
대치어		노루,여우,개 (새,물고기를 제외한 동물)		꼬리	

<표 3> 의미 유사성 오류의 예

(예 2)는 '꼬리/꽁지'의 의미 유사성으로 인해 오용되는 경우이다. '꼬리'는 새나 물고기를 제외한 동물의 꽁무니나 몸뚱이의 끝을 의미한다. '꽁지'는 새나 물고기의 꽁무니를 의미한다. '꼬리'와 '꽁지'는 그 의미는 유사하지만 쓰이는 주체가 다르다.

o. 발음 유사성으로 인한 오류

발음은 비슷하나 의미가 다른 단어끼리 오용되는 경우이다.

(예 3) 긴 인간의 역사를 두고 이어져 내려오던 가치의 기준이 서로 엉켜 버린 실오리처럼 뒤범벅이되고,

지배 관계		지배소		의존소	
문장	긴 인간의 역사를 두고 이어져 내려오던 가치의 기준이 서로	엉켜	버린	실오리처럼	뒤범벅이되고
연관 관계를 가지는 단어		엉키다		실, 머리카락	
대치어		엉기다		피, 기름	
대치어		얽히다		싸움, 회담, 사건, 일	

〈표 4〉 발음 유사성으로 인한 오류의 예

(예 3)은 '엉기다/엉키다/얽히다'의 발음 및 의미 유사성으로 인해 오용되는 경우이다. '엉기다'는 응고(凝固)의 의미이고, '엉키다'와 '얽히다'는 이리저리 섞여서 묶여진 것을 의미한다[3]. '엉기다'와 '엉키다/얽히다'는 의미는 전혀 다르지만 발음이 유사하기 때문에 빈번히 오류를 발생시킨다. 한편 '엉키다'는 '실/머리카락'등의 명사와 함께 사용되고, '얽히다'는 '일/사건/회담/싸움'등의 명사와 함께 사용된다. 이 때 '엉키다'와 '얽히다'는 의미 유사성으로 인한 오류이다.

o. '부사-서술어'의 관용구 사용 오류

'부사어-서술어'로 구성된 관용구에서 '부사어'에 대한 '서술어'의 의미 차이로 인한 오류이다.

(예 4) 너는 결코 시키는대로 해야 한다.

지배 관계		지배소		의존소
문장	너는	결코	시키는대로	해야 한다
연관 관계를 가지는 단어		결코		'아니다, 말다, 못하다 않다' 등의 부정어

〈표 5〉 '부사-서술어'의 관용구 사용 오류의 예

'결코'는 부정하는 말과 함께 쓰여 '절대로'의 뜻을 가진다. (예 4)는 서술어인 '해야 한다'가 긍정이므로 '결코'와 의미상의 차이가 생긴다.

o. 조사 오류

동사의 종류와 조사와의 관계에 의한 오류이다.

(예 5) ㈀ 먹고 사는 일이 나를 얽매인다.
㈁ 낡은 전통에 얽매서 헤어나지 못한다.

지배 관계		의존소	지배소
문장	먹고 사는 일이	나를	얽매인다
연관 관계		-에	얽매이다
대치어		-을/를/ㄹ	얽매다

〈표 6〉 조사 오류의 예

(예 5)의 ㈀은 타동사 '얽매다'를 써야 하는 경우이고 ㈁은 자동사 '얽매이다'를 써야 하는 경우이다. '얽매다/얽매이다'는 연관 관계의 명사가 너무나 많으므로 명사와의 연관 관계보다는 조사와의 제약이 현실성이 있다. '얽매다/얽매이다'는 '-을/를/ㄹ 얽매다', '-에 얽매이다'의 형태로 사용된다.

o. 용언대 용언의 의미적 지배 오류

본용언과 보조용언 간의 의미적 차이에 의한 오류이다.

(예 6) 벗이 동행하기를 청한다면 함께 가 오라.

지배 관계		지배소	의존소
문장	벗이 동행하기를 청한다면 함께	가	오라
연관 관계를 가지는 단어 (긍정적)		가다	버리다, 보다, 주다, 지다, 달다
연관 관계를 가지는 단어 (부정적)		가다	가다, 오다, 나가다, 나오다, 놓다, 대다, 두다, 나다, 넣다

〈표 7〉 용언대 용언의 의미적 지배 오류의 예

용언대 용언 오류는 빈번히 범하지는 않는다. 이 오류를 철자 검사기에서 처리하기 위해서는 기본적인 의미 정보에 의해 보조용언을 다시 분류하는 등의 노력이 필요하므로, 교열 시스템에서 처리하는 것이 타당하다. (예 6)은 긍정적 연관 관계와 부정적 연관 관계를 모두 제시한 예이다. 본용언 '가다' 뒤에 보조용언 '오다'가 이어진 '가오다'는 우리말에서 전혀 쓰이지 않는 단어이다. 반면에 '오다' 뒤에 보조용언 '가다'가 이어진 '오가다'는 우리말에서 빈번히 쓰이고 있는 올바른 말이다. 이 경우 '오다'를 지배소로 하는 연관 관계에서는 '가다'는 긍정적 연관 관계로 분류되어야 한다.

o. 의미 중복 오류

우리말은 한자어와 순수한 한국어를 복합하여 사용한다. 한자는 의미 문자이므로 한 자어 마다 의미를 내포하고 있다. 한글은 소리 문자이므로 단어가 되어야만 그 의미가 부여된다. 의미 중복 오류는 한자와 한글의 이러한 차이로 인해 발생한다.

(예 7) ㉠ 길을 가다 봉변을 당했다.

지배 관계		의존소	지배소
문장	길을 가다	봉변(逢變)을	당했다

〈표 8〉 의미 중복 오류의 예1

㉡ 역전 앞에서 만나자.

지배 관계	지배소	의존소	
문장	역전(驛前)	앞에서	만나자

〈표 9〉 의미 중복 오류의 예2

(예 7)의 ㉠에서 '봉변'에서 봉(逢)은 이미 '당한다'는 의미를 내포하고 있으며, ㉡의 '역전'에서 '전(前)'은 '앞'이라는 의미를 내포하고 있다.

o. 존칭 오류

한국어는 다른 어느 언어보다 존칭법이 발달했다. 평서형의 문장에서 쓰는 단어가 존칭형의 문장에서 바뀌는 경우가 있으며, 용언에 존칭 선어말 어미가 결합하여 존칭형 문장의 서술어가 된다. 또한 존칭 선어말 어미와 결합되지 않고 이미 굳어진 형태

의 존칭형 서술어가 있다. 이러한 존칭법의 어려움으로 인해 좋지 않은 존칭형 문장이 쓰이는 경우가 빈번하다.

(예 8) 아버지께서 밥을 먹으신다.

지배 관계	지배소	의존소	의존소	의존소
문장	아버지	께서	밥을	먹으신다
연관 관계를 가지는 단어	아버지	께서	진지, 식사	잡수시다, 드시다

<표 10> 존칭 오류의 예

(예 8)은 하나의 지배소가 여러 개의 의존소를 지배한다. 존칭형 문장에서는 주로 존칭의 주체가 지배소가 되고 지배소를 제외한 단어들이 의존소가 된다. (예 8)의 조사 '께서'는 올바른 표현이고, 서술어 '먹으신다'와 같은 표현은 좋지 않다. '밥'은 '진지'나 '식사'로 바꿔어야 하고, '먹으신다'는 '잡수신다'나 '드신다'로 바꿔어야만 한다.

o. 띄어쓰기 오류

철자 검사기에서 띄어쓰기 오류가 가능한 경우는 떨어진 단어 중 어느 하나가 철자 검사에서 오류 어절로 판단될 때이다. 떨어진 두 어절 모두 맞는 어절로 판단될 경우에는 현재의 철자 검사기로는 처리가 불가능하다[5]. 예외적으로 부산대 철자 검사기는 (예 9)와 같은 '접미사/불완전 명사/어미'의 혼용 오류를 처리할 수 있지만 극히 제한된 경우에 한하고 있다.

(예 9) ㉠ 친구뿐이다.(o) : 접미사
㉡ 친구 뿐이다.(x)
㉢ 잘 뿐이다.(o) : 불완전 명사

(예 10) ㉠ 먹을 수록(x)
㉡ 먹을수록(o)

지배 관계	의존소	지배소
문장	먹을	수록
용언-어미 지배 (긍정적)	ㄹ 종성 관형형 어미	수록[어미]

<표 11> 띄어쓰기 오류의 예

(예 10)의 ㉠은 철자 검사기에서 '먹을'과 '수록(手錄, 收錄)'으로 처리되므로 오류를 알 수 없다. 그러나 '종성 ㄹ인 용언'과 '수록'은 붙인다는 규칙을 규칙 베이스에 추가한다면 한국어 교열 시스템에서 처리 가능하다.

o. 기타

이상에서 유형 분류한 오류 외에도 문장 부호 오용, 유사 한자어 오용, 속담 및 격언의 오용 등이 있다. 본 논문에서는 이와 같은 오류는 빈번하지 않다고 보고, 이에 대한 처리는 제외하였다.

3.오류 유형의 정형화

오류 유형	오류 단어의 관계	연관 관계
발음 유사성 오류	명사와 용언 간의 오류	긍정,부정
의미 유사성 오류	명사와 명사 간의 오류	긍정,부정
부사-서술어 사용 오류	부사와 서술어 간의 오류	긍정
조사 오류	동사와 조사 간의 오류	긍정,부정
용언대 용언 오류	본용언 과 보조용언 간의 오류	긍정,부정
의미 중복 오류	명사와 서술어, 혹은 명사와 명사 간의 오류	부정
존칭 오류	명사와 서술어, 혹은 명사와 명사 간의 오류	긍정,부정
띄어쓰기 오류	용언과 어미 간의 오류	긍정,부정

<표 12> 오류 유형의 분류

Ⅲ.한국어 교열 시스템을 위한 규칙 베이스의 구조

1.규칙 베이스의 구조

본 논문에서 제안하는 오류 유형에 대해 각기 다른 규칙 베이스를 구성할 수 있다. 그러나 이 방법은 처리가 일관적이지 못하며, 규칙의 추가/삭제가 복잡하다는 단점이 있다. 그러므로 모든 오류 유형의 다양한 특징을 표현할 수 있는 전형적인 구조가 필요하다. 또한 교열 시스템은 한 문장을 여러 번 분석하게 될 가능성이 있으므로 처리 속도가 문제시 된다. 이 경우 규칙 베이스 탐색 시간 역시 시간 지연의 원인이 되므로 직접적인 규칙 베이스 탐색으로 시간 지연을 최소화 시킬 수 있는 구조가 필요하다. 이와 같은 요구 사항들을 고려하여, 다음 그림과 같은 구조의 규칙 베이스를 제안하여 한국어 교열 시스템의 구현에 응용한다.

```
# Sub-rule N
  (( Governor 1 (PR ( W1-P1, I) ( W1-P2, I) ... ( W1-Pn, I))
                (NR ( W1-N1, I) ( W1-N2, I) ... ( W1-Nm, I)))
   ( Governor 2 (PR ( W2-P1, I) ( W2-P2, I) ... ( W2-Pn, I))
                (NR ( W2-N1, I) ( W2-N2, I) ... ( W2-Nm, I)))
       ...
   ( Governor k ...)
  )
```

* Sub-rule N : Sub-rule number
Governor k : Sub-rule N에서의 k번째 지배소
PR : 긍정적 관계, NR : 부정적 관계
W_{k-Pn} : k번째 지배소와 PR 관계를 가지는 의존소
W_{k-Nm} : k번째 지배소와 NR 관계를 가지는 의존소
I : 격정보(명사), 어미정보(용언)

<그림 1> 규칙 베이스의 구조

규칙 베이스는 오류 유형을 Super-rule로 명시하고, 각 오류 유형에 대한 세부적인 대치어쌍을 각각 하나의 Sub-rule로 분류하여 명시한다. 즉, 규칙 베이스의 Rule-number는 계층적으로 구성되어 있다. 오류의 Sub-rule은 인덱스에 의해 직접적으로 접근되므로 규칙 베이스 접근 시간 지연은 최소화 된다. Governor는 세부적인 대치어들이다. Governor는 지배 관계에서의 지배소이고 W(Word)는 의존소이다. 지배소와 긍정적인 연관 관계에 있는 단어들은 PR field에 나열된다. 즉, $W_{1\text{-}P1}$부터 $W_{1\text{-}Pn}$까지의 단어가 Governor i와 문장에서 함께 사용될 경우 그 문장은 의미적으로 옳다고 볼 수 있게 된다. 지배소와 부정적인 연관 관계에 있는 단어들은 NR field에 나열된다. NR은 PR의 반대 개념이다. I(Information) field는 오류 유형별로 각 의존소의 특별한 정보를 보유한다. 앞 장에서 분류한 오류 유형에서 지배소는 명사이거나 용언이고, 의존소는 '명사+조사' 혹은 '어간+어미'의 유형이라는 유사점이 있다. 그러므로 위의 구조에서 W field를 용언의 어간 혹은 명사로 하고, I field를 어미 정보 혹은 조사 정보로 하면 이와 같은 유형들을 일관적으로 표현할 수 있다. 어미 정보는 앞 장의 (예 10)과 같은 경우에 필요하며, 조사 격정보는 (예 5)와 같은 조사 오류나, 의미 및 발음 유사성 오류에서 연관 관계 명사의 제약을 강화하기 위해 필요하다.

어떤 문장에서 PR 관계의 단어와 NR 관계의 단어가 모두 발견될 경우 현재 문장의 오류를 결정하는 데는 어느 단어가 빈번히 사용되느냐가 관건이다. 본 규칙 베이스에서는 가중치가 높은 단어를 앞에 위치 시킴으로써 자연스럽게 위치 정보로써 가중치를 표현하도록 한다. I field는 의존소가 명사이면 조사의 격을 의미하며, 용언이면 어미의 종류를 의미한다. 조사는 <표 13>과 같이 분류한다. 이 때 보조사는 문장 분석을 한다 할 지라도 격 결정이 어려우므로 정보 분류에서 제외한다. 용언의 어미는 표14)와 같이 분류한다.

분 류	조 사
주격	은/는/이/가,께서
서술격	이다
목적격	을/를/ㄹ
보격	이/가
관형격	의
부사격	에,에서,에게,(으)로
호격	아/야,(이)시여,(이)여
접속 조사	와/과,하고,(이)며

<표 13> 조사의 분류

분 류	어 미
관형형 어미	ㄴ/은/는
연결형 어미	아다/어다, 아야/어야
종결형 어미	ㄴ가,ㄴ고,ㄴ걸
명사형 전성 어미	ㅁ/음/기

<표 14> 어미의 분류

2.규칙 베이스의 예

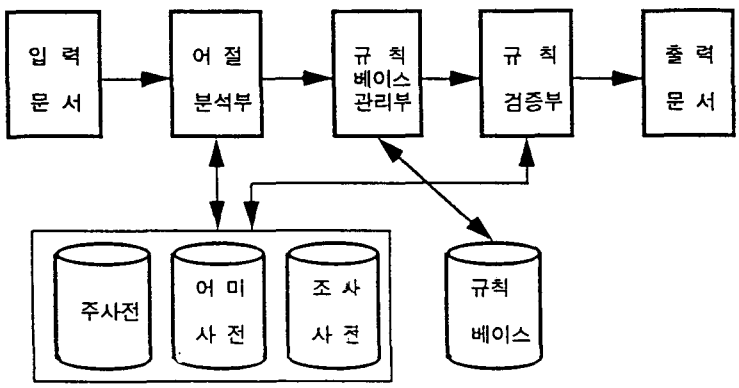
```
#.2.1
  ((엉기다 (PR (피,주격) (기름,주격) ...)
           (NR (실,주격) (머리카락,주격) ... ))
   (엉키다 (PR (실,주격) (머리카락,주격) ...)
           (NR (피,주격) (기름,주격) ... ))
   (얽히다 (PR (싸움,주격) (회담,주격) (사건,주격) (일,주격) ...)
           (NR (실,주격) (머리카락,주격) ... )))
# 3.1
  ((얽매이다 (PR (DC, 부사격[처소]))
             (NR (DC, 목적격)))
   (얽매다   (PR (DC, 목적격))
             (NR (DC, 부사격[처소])))
# 4.1
  ((가다 (PR (버리다,I) (보다,I) (주다,I) (지다,I) (달다,I) ... )
         (NR (가다,I) (오다,I) (나가다,I) (나오다,I) ... )))
```

<그림 2> 규칙 베이스의 예

위 그림은 앞 장의 오류 유형 중 (예3,5,6)의 규칙 베이스를 보인 것이다. Sub-rule 2.1은 발음 유사성 오류에 대한 규칙 베이스이다. '엉기다'에서 '피/기름'에 주격조사가 결합된 형태는 올바른 문장이다. 일반인들은 '엉기다'를 '실/머리카락'과 함께 사용하는 오류를 빈번히 범한다. 그러므로 NR field에서는 오류를 빈번히 범하는 단어만 명시한다. NR과 PR에 명시되지 않은 단어는 고려하지 않으므로 그러한 단어와 '엉기다'가 함께 사용되면 올바른 문장으로 본다. Sub-rule 3.1은 조사 오류에 대한 규칙 베이스이다. 이 규칙은 명사는 고려하지 않고 격조사만을 고려하므로 단어 field는 DC(Don't Care)로 두어 고려에서 제외한다. Sub-rule 4.1은 용언대 용언의 의미 지배 오류의 규칙 베이스이다. 이 경우 용언의 어미 정보가 필요하지는 않으므로 I field는 사용되지 않는다.

Ⅳ.한국어 교열 시스템의 구현

1.한국어 교열 시스템의 구조



<그림 3> 교열 시스템의 구조

어절 분석부는 한국어 교열 시스템의 인터페이스 부분이다. 이 부분에서는 입력 문서로부터 한 문장을 추출하여 어절 단위의 간단한 형태소 분석을 행한다. 간단한 형태소 분석이라 함은 어절 내의 '명사+(명사)+조사' 혹은 '본용언+(보조용언)+어미'의 형태를 분석함을 의미한다. 현재 구축된 규칙 베이스에서 필요로 하는 모든 정보는 이 정도의 최소한의 형태소 분석으로 추출해 낼 수 있다. 어절 분석부에서는 주사전과 어미,조사 사전을 참조한다. 교열 시스템이 목적으로 하는 의미적 오류의 가능성이 있는 단어일 경우 주사전 검색 후 반환된 정보는 단어의 품사적 정보와 규칙 베이스의 Rule-number이다. 어절 분석부는 현재 문장과 사전 정보를 다음의 규칙 베이스 관리부로 넘긴다. 다음은 어절 분석부의 처리 알고리즘이다.

```
문서의 끝이 아닌 동안
   문서에서 한 문장 추출
   문장의 끝이 아닌 동안
      문장에서 한 단어 추출
      주사전에서 단어 검색
      주사전에 있고 오류 가능 단어이면
         정보가 명사이면 복합명사,조사 처리
         정보가 용언이면 보조용언,어미 처리
         명사/용언 처리를 통과하면
            규칙 베이스 관리부 호출
      문장에서 다음 단어 추출
   문서에서 다음 문장 추출
```

<그림 4> 어절 분석부 처리 알고리즘

규칙 베이스 관리부는 어절 분석부에서 넘어온 Rule-number로써 규칙 베이스를 직접 접근한다. 규칙 베이스 관리부는 오류를 발생시킬 가능성이 있는 것으로 판단된 현재 단어에 대한 규칙 전체를 규칙 베이스에서 메모리로 가져온다. 이 때 규칙 베이스의 규칙들은 PR과 NR로 분리되어 PR 검증부와 NR 검증부로 나뉘어 유지된다.

```
규칙 베이스 인덱스 접근
인덱스 값으로 규칙 베이스 접근
Sub-rule번호 비교
   일치하면
      후보 단어 설정
         PR 관계이면 PR 검증부로 의존소 저장
         NR 관계이면 NR 검증부로 의존소 저장
```

<그림 5> 규칙 베이스 접근부 알고리즘

규칙 검증부는 규칙 베이스에서 가져온 규칙에 의해 현재 문장에 대해 필요한 부분만 재분석한다. PR 및 NR 검증부는 저장된 규칙 정보와 의존소들을 문장 재분석을 통해 적용해보고, 일치되면 가중치를 변화시킨다. 문장의 분석이 완료되면 그 가중치에 의해 문장의 옳고 그름을 판단한다. 그른 문장의 경우 한국어 교열 시스템은 PR및 NR에 저장된 규칙을 형식화(Formatted)하여 오류 이유와 함께 제시한다. NR은 PR과 가중치의 변화를 제외하고는 처리 과정이 같으므로 알고리즘을 생략했다. PR검증부에서는 하나의 문장 내에서 서술어에서 서술어까지만 분석한다. 그 이유는 분석 범위가 절 단위를 넘어설 경우는 격조사나 어미의 종류 등을 예측할 수 없으므로 규칙 베이스의 규칙을 적용하는 것이 적합하지 않기 때문이다.

```
PR검증 수행
NR검증 수행
최종 판단
옳지 않은 문장이면
    형식화된 출력
```

〈그림 5〉 규칙 검증부 알고리즘

```
오류 가능 어절이 관형형이면
   처리방향 좌에서 우로 설정
아니면
   처리방향 우에서 좌로 설정

모든 후보 지배소에 대해
   문장에서의 현재 위치부터 문장 끝까지
      문장에서 한 어절 추출
      Super-rule구분
         의미 유사 오류   :
         조사 오류        : 명사+(명사)+조사 중점 검사

         발음 유사 오류   :
         부사-서술어 오류 :
         용언대 용언 오류 :
         띄어쓰기 오류    : 본용언+(보조용언)+어미 중점 검사

         존칭 오류        :
         의미 중복 오류   : 명사+(명사)+조사 검사
                            본용언+(보조용언)+어미 검사

      이전의 서술어에 대한 의존소가 존재하지 않고
      새로운 서술어 등장하면
         현재 문장을 올바른 문장으로 판단
      아니면
         다음 어절 추출
```

〈그림 6〉 PR검증부 처리 알고리즘

특히, 지배소가 용언이고 의존소가 명사일 경우는 검사 방향이 문제시 된다. 즉, 오류 가능성이 있는 어절이 관형형일 경우 그 처리 방향이 좌에서 우가 되어야 하고, 관형형이 아니면 우에서 좌가 되어야 한다. 예를 들어, '엉킨 피가'의 경우는 지배소인 '엉킨'이 관형형 어미로 끝나므로 바로 다음 어절인 '피가'를 검사한다. 일반적으로 관형형 어미 뒤에는 수식되는 명사가 가까이 있으므로 이 경우는 다음 번에 나타나는 용언이나 문장의 끝까지의 명사를 검사한다. '피가 엉키다'의 경우는 지배소 '엉키다'가 서술형이므로 반대 방향의 어절인 '피가'를 검사한다. 이 경우 역시 새로운 용언이 나타나거나 문장의 처음이 되면 검사를 중단한다. 그러나 이 방법은 완벽한 관형형 분석이 선행되지 않으면 반대 방향 만을 검사해 버릴 수 있다는 문제점을 안고 있다. 이 부분은 향후 개선될 여지가 있다.

주사전 및 어미/조사 사전은 부산대학교에서 개발된 변형된 트라이(Trie) 구조를 사용한다. 주사전은 규칙 베이스에 수록된 단어뿐만 아니라 어절 분석을 위한 단어 및 정보를 수록하고 있다. 규칙 베이스에 지배소로 수록된 단어는 품사 정보외에 추가의 정보로 규칙 번호를 가진다. 어미와 조사 사전은 앞에서 살펴본 어미 분류와 조사 분류 정보 및 어절에서의 어미와 조사 처리를 위한 제약 조건들을 보유한다.

V.한국어 교열 시스템의 처리 예

o.발음 유사성으로 인한 오류

문장 : 머릿속에서 온갖 생각들이 엉기고 있단다

제시어 : 엉기다
제시어 : 엉키다
제시어 : 얽히다

선택구문: 얽히다 + { 생각 }

o.의미 유사성으로 인한 오류 예

문장 : 고늠의 개공지가 빠져라 도망치는구나

제시어 : 꼬리
제시어 : 꽁지

선택구문: 꼬리 + { 개 }

o.조사 오류

문장 : 자고로 공직에 몸담은 사람은 사사로운 일에 얽매면 안된다

제시어 : 얽매다
제시어 : 얽매이다

선택구문: 얽매이다 + { 일에 ¦ 공직에 }

o.부사-서술어 오류

문장 : 그것만은 결코 알아주십시오

선택구문: 결코 + { 않다 ¦ 마다 ¦ 아니다 ¦ 못하다 }

o.용언대 용언의 의미적 오류

문장 : 일단 한번 와 오면 알것 아냐

제시어 : 오다

제거구문: 오다 + { 오다 }

o.의미 중복 오류

문장 : 오늘 오후 역전 앞에서 만나는 거다

제시어 : 역전

제거구문: 역전 + { 앞 }

o.존칭 오류

문장 : 아버지께서 진지를 먹으신다

제시어 : 아버지

제거구문: 아버지 + { 먹으시다 }

o.띄어쓰기 오류

문장 : 먹으면 먹을 수록 그 맛이 더해지는 구나

제시어 : 수록

선택구문: 수록 + { ㄹ ¦ 을 }

VI.결론

본 논문에서는 실데이타의 분석을 통해 단어 사이의 오류 유형을 분류하였고 이러한 오류를 처리하는 교열 시스템을 제안하였다. 교열 시스템은 통사 분석에서 쓰이는 지배 관계를 변형시켰으며, 의미 처리를 위하여 단어 사이의 연관 관계를 규칙 베이스화 했다. 또한 철자 검사기가 처리할 수 없는 단어 사이의 의미적 오류를 최소한의 형태소 분석만으로 처리했다. 본 시스템은 완벽한 통사 분석이나 의미 분석 없이 실데이타의 분석을 통해 구축된 규칙 베이스 만으로도, 범하기 쉬운 단어 사이의 의미적 오류를 검사하고 교정할 수 있음을 보여준다. 즉 한국어 문체 검사기(Korean Style Checker)의 원형으로서 그 의의를 가진다.

본 논문에서 제시한 교열 시스템은 아직 완벽한 문장 단위의 의미적 오류를 처리할 수는 없다. 그러나 규칙 베이스의 정보가 충분히 많아지고, 파싱등의 문장 분석 기법

에 이 정보를 이용하면 문장 단위의 의미적 오류 처리가 상당히 개선될 것이다. 향후 과제로서 일반인들의 의미적 오류 유형의 더욱 면밀한 조사가 필요하며, 그것을 기반으로한 규칙 베이스의 지속적인 개선이 필요하다.

참고 문헌

[1]이 영식 및 권 혁철,"한국어 철자 오류 교정 시스템", 1993년도 춘계 학술발표논문집, 한국정보과학회, 인공지능연구회, pp. 178-187, 1993

[2]김 민정 및 권 혁철 "한국어 형태소 분석에서의 수사 처리", 제3회 한글 및 한국어 정보 처리 학술발표논문집, 한국인지과학회, pp.813-862, 1991

[3]미 승우, "새맞춤법과 교정의 실제", 어문각, 1988

[4]Hyuk-Chul Kwon, Aesun Yoon, "Unification-Based Dependency Parsing of Governor-Final Languages", Proc. of Second International Workshop on Parsing Techonolgy, Cancun, Mexico, pp.182-192, 1991

[5]옥 철영, "한.영 기계번역을 위한 구 단위 변환 사전",서울대학교 박사학위 논문, 1992

[6]Masaki YAMASHINA, "Collocational Analysis in Japanese Text Input", Coling budapest '88 Vol II, 1988

[7]Paola Velardi, "How to Encode Semantic Knowledge:A Method for Meaning Representation and Computer-Aided Acquisition", Computational Linguistics, Vol.17. No.2, 1992

[8]박 나리,김 영택,"Collocation 정보에 기반한 한.영 기계번역 사정의 구성", 1991년도 추계 학술발표논문집, 한국정보과학회, pp.833-836, 1991

[9]박 인환, "실용국어 중심의 바른 우리말", 세계일보, pp.15-155, 1991

[10]"중앙일보사 Style Book", 중앙일보