비대칭 관계에 기반한 지식베이스와 문법 검사기 구현*

강미영[0] 임성신 권혁철
부산 대 학 교 전자자계산학 과 인공지능 연구실
{kmyoung, sslim, hckwon}@pusan.ac.kr

Implementation of Knowledgebase and Grammar Checker Based on Asymmetric Relationship

Mi-Young Kang[0] Sung-Shin Lim Hyuk-Chul Kwon
AI Lab. Dept. of Computer Science, Pusan University

요 약

이 논문은 한국어 문서 교정을 위해 부산대학교 인공지능 연구실에서 개발되고 있는 맞춤법 및 문법 검사기와 그 지식베이스 구축에 대해 알아본다. 비대칭 관계(asymmetric relation) 설정과 더불어 개발된 문법 검사기는 한 문장의 임의의 두 요소 간의 비대칭 관계를 토대로 ① 잠재적 지배관계 개념을 설정하고 ② 부분 문장 분석 방법을 사용한다. 이런 요소들을 효율적으로 운영하는 시스템을 구현하기 위해서는 한국어에 대한 이해와 경험적 자료를 적절히 조화시킨 지식베이스 구축이 필수적이다. 이 논문은 그 선수(先手) 작업으로, 한국어 문서의 오류들을 관찰하고, 그로부터 추출한 언어적이거나 언어 외적인 요소들로부터 일반적 패턴을 뽑아내어 사용자의 기대를 만족시키기 위한 적절한 규칙 지식베이스를 구축하고 문법 검사기의 성능 향상을 위해 문장 분석 출발점과 분석방향 그리고 분석범위를 한정한다.

1. 서론

한국어 처리는 1970년대 말부터 시작되었다. 초기 연구는 주로 규칙과 지식베이스 기반 연구로, 다룰 수 있는 범위가 한정되어 있었다. 한국어는 교착어적인 경향이 강하여 형태소의 기능이 비교적 잘 분리되어 있고 각각의 형태소로 잘 자를 수 있다고는 하더라도 어간을 분리하는데 많은 어려움이 따른다. [11] 또 다른 한국어에 고유한 어려움으로 띄어쓰기를 들 수 있는데 의미·통사적인 면에 영향을 끼치며, 때로는 선택적이라는 문제점을 보인다. 이와 같은 형태론적 복잡성이 영어 처리보다 한국어 처리시에 낮은 시스템 성능을 보이는 원인이 되어왔다. 이러한 단점을 보완하기 위하여 1990년 초에 코퍼스 기반으로 한 태깅 시스템이 만들어지고 형태소 분석기가 개발되었으며[ETRI, KAIST], 1990년 중반에는 확률적 방법이 말뭉치 기반연구와 더불어 도입되었다.[1]

확률과 규칙에 기반한 한국어 문법 검사기가 부산대학교 인공지능 연구실에 의해 구현되어[11] 규칙이 먼저 적용되고 만약 동일한 값을 가진 두 가지 요소가 있을 때 코퍼스 연구를 통해 미리 얻어진 통계적인 가치에 따라 통계적인 선호도를 부여함으로써 선택 한다.

완전 문장 분석은 한국어 문서 분석을 하는 동안 발견되는 여러 종류의 오류를 이상적으로 처리할 수 있다. 그러나 이러한 시스템은 시간 낭비가 크다는 단점뿐만 아니라, 한국어의 유형론적인 특성에 대처하기 위해서는 많은 추가 규칙을 설정해야 한다는 단점이 있다. 이러한 문제점을 해결하기 위하여 부분 문장 분석 기법이 도입되었다.[3], [4], [5], [10]

이 논문에서 소개하는 한국어 맞춤법 및 문법 검사기(버전 2.0)는 잠재적인 지배소를 설정하여 부분 문장 분석 사용한다. 즉, 한 문서 내에서 오류나 중의성이 발견되면 주어진 지배소가 앞서는 어절이

* 이 논문은 과학 기술부(한국과학기술기획평가원)의 국가 지정 연구실 지원으로 이루어진 것임.

[1] 90년대 한국어 철자 및 문법 검사기에 관한 연구로는 본 연구실 외에도 과학 기술원[강재우 90], [송춘환 90] 등, 서울대[박종만 90], [강승식 2002] 등, 연세대[박영환 92], [이병훈 93] 등이 있는데, 현재는 여러 연구소와 개발 업체별로 연구가 활발하다.

나 뒤따르는 어절과 함께 나타날 수 있는지 검사한다. 이러한 검사 모듈을 최적화하기 위하여 의존 문법에 기반하여 문장의 두 요소간의 비대칭 관계를 설정한다. 일반적인 의존 문법은 문장 성분들간의 결합 순위와 지배관계를 연구하며 한국어를 주로 동사가 최상의 위치에 자리잡고 있는 언어로 분석 한다. 이 때 상위에 있는 성분을 지배소라 부르고 이것에 의해 지배되는 성분을 피지배소라고 부른다. 본 문법 검사기에서의 지배소는 일반적 의존 문법에서 일컫는 의미로 받아들여진 것이 아니고 시스템이 오류 어절이나 중의적인 어절로 지적할 가능성이 높은 어절이나 구의 패턴을 말한다. 이러한 비대칭 관계에 기반한 부분 문장 분석 방법과 잠재적 지배관계의 개념은 ① 부분 문장 분석 실행 조건과 ② 분석 범위와 ③ 분석 방향을 제공함으로써 분석 시간 절약과 시스템의 정확도를 보장해준다.

이 논문의 구성은 다음과 같다. 2장에서는 문장 분석을 하는 동안 만날 수 있는 오류들을 사용자들의 가능한 언어학적 혼동과 연관하여 기술한다. 3장에서는 비대칭 관계에 기반하여 구축된 부분 문장 분석을 소개한다. 4장에서 한국어의 특성에 기반한 규칙과 단서(clue)를 알아보고 규칙처리 모듈의 운영에 대해 소개한다. 마지막 장에서는 남은 문제를 알아보고 향후 연구의 방향을 살펴본다.

2. 언어 사용 오류 분석

한국어 사용자는 여러 언어학적인 단계, 즉 음운론적, 어휘적, 형태론적, 의미론적, 통사론적 단계 등에서 혼동을 일으킬 수 있다. 여기에 또한 관용어 표현과 외래어 표기에 대한 무지 및 혼동을 포함 시킬 수 있는데 이러한 영향은 한국어 문서를 분석하는 과정에서 발견되는 여러 오류들의 유형을 통해 드러난다.

약 18,000,000 어절을 포함하는 1년치 신문 데이터를 본 연구실의 맞춤법 및 문법 검사기(2.0)로 돌려서 분석한 결과 약 1,254,505 어절의 오류가 발견되었다. 본 시스템의 미등록어와 관련된 오류는 149,964 어절로써 전체 오류의 11.95%에 해당한다. 이중 약 90%가 고유 명사이다. 이러한 미등록어를 그 유형별로 분석해 보면 다음과 같다.

대치어 유무	미등록어	빈도수
대치어 없음	의미 자질 {+ human}	20.50%
	의미 자질 {- human}	46.58%
	외래어	0.79%
대치어 있음	의미 자질 {+ human}	13.01%
	의미 자질 {- human}	18.91%
	외래어	0.21%

[표 1] 미등록어 유형과 빈도수

다음은 미등록어 관련 오류를 제외한 1104541어절에 대한 오류 유형과 그 출현 빈도수다.

오류 유형	오류 어절 수	빈도수
의미·문체 오류	48837	4.42%
띄어 쓰기 오류	263372	23.84%
유사 발음 관련 오류	126294	11.43%
철자 오류	325799	29.50%
어휘 조합 오류	8569	0.78%
동사 활용 오류	303	0.03%
표준어 사용 오류	73275	6.63%
순화 용어 오류	44281	4.01%
외래어 표기 오류	8015	0.73%
문장 부호 오류	205794	18.63%
관용어 오류	2	0.00%
전체 오류 어절 수	1104541	100.00%

[표 2] 오류 유형과 빈도수

위 테이블의 오류 유형 중에서 의미·문체 오류는 통사적 구조의 잘못된 표현에 의해 의미 문체적으로 받아들여질 수 없는 문장을 만드는 오류를 분류해 놓은 것이다. 띄어쓰기 오류 중에서 12%의 어절들은 품사 혼동에서 온 것이라 할 수 있다. 한편 유사 발음 및 음운론적 현상에 따른 혼동에서 빚어진 많은 오류들이 있으며 사용자의 입력 오류나 철자 혼동에서 비롯한 철자 오류가 있다. 어휘 조합이나 동사 활용 오류는 형태론적 어휘 형성 조건에 대한 혼동과 관련 있으며 표준어 사용 오류나, 순화용어 사용 오류 같은 문체 오류나 관용어 표현 오류, 외래어 표기 오류와 같은 규범 문법관련 오류가 있다. 이와 같이 한 문서를 분석할 때 발견되는 오류들에 대한 처리는, 해당 문서를 만든 사람의 언어학적 능력과 사용에 대한 분석 및 이해를 통하여 보다 더 효과를 거둘 수 있는데, 이는 각각의 오류 유형에 대처할 수 있는 적합한 지식베이스를 구축할 수 있기 때문이다.

2.1 통사적 구조 혼동

한국어 문서의 많은 오류들이 잘못된 통사적 구조에서 연유한다. 이러한 통사적 구조은 동사의 하위범주화에 의해 결정된다. 동사는 특정한 논항과 더불어 나타날 수 있다. 이와 더불어 조사들은 명사들의 통사적 기능을 한정한다. 주격(명격)(이/가), 대격(을/를), 여격(에게), 처소격(에), 소유격(의), 주제어 표지 (은/는) 등[2]이 대표적인 격

[2] 대각선 왼쪽의 첫번째 이형태(異形態)는 자음으로 끝나는 명사 다음에 나타나는 것이고 두번째 이형태는 모음으로 끝나는 명사 다음에 나타나는 조사이다.

표지 조사들이다. 이러한 조사가 붙은 명사들은 주어(←주격), 목적어(←대격), 간접 목적어(←여격)과 같은 통사적 기능을 담당한다. 또한 처소격은 일반적으로 장소나 목적 등을 표시하기 위해 사용되며 소유격은 명사의 보어 역할을 담당하게 된다. 따라서 동사와 조사의 일치는 문장의 문법성을 결정하게 된다. 이에 따라 자동사는 항상 주격 조사가 붙어 있는 논항을 취하는 반면 타동사는 주격을 취하는 논항과 대격을 취하는 논항을 취하게 된다. 이중타동사는 세 개의 논항을 취하는 동사로서 여격이 세번째 논항으로 할당되게 된다.[3]

(1) 우리-가#적-을#이기-고
명사-주격조사 /명사-대격조사 /타동사-접속어미
(1') * 우리-가#적-에게#이기-고
명사-주격조사 /명사-여격조사 /타동사-접속어미
(2) 내-가#때리-ㄴ#동생-이
대명사-주격조사 /동사-관계절 어미 /명사-주격조사
(2') * 나-의# 때리-ㄴ#동생-이
대명사-소유격조사 /동사-관계절 어미 /명사-주격조사

(1)의 동사 '이기-' 는 타동사이므로 (1')의 여격 조사가 붙은 논항 '적-에게' 와는 양립할 수 없다. 또한 (2')의 소유격 조사가 붙은 논항 '나-의' 는 명사 앞에서 명사의 보어 역할을 수행해야 하는데 동사 앞에 옴으로써 비문법적인 문장을 만들어내게 된다.

2.2 품사 표현 오류

많은 한국인들은 품사 표현 오류를 범하며 이러한 오류는 상당 부분 띄어쓰기 오류를 통해 드러난다. 한국어 규범 문법은 다음과 같이 품사에 따라 띄어쓰기를 달리 정하여 조사나 접두사를 대범주 어절에 붙여 쓰고 관형사는 그가 수식하는 명사와 띄어쓰기를 하고 부사는 앞뒤를 다 띄어 쓰게 규정하고 있다. 그들의 통사적인 특성은 띄어쓰기를 통해서 드러난다.

(3) 조사와 부사 사이의 혼동
(3') 현실-보다#이상-을#쫓고
(3") 현실#보다#이상-을#쫓고
(4) 관형사와 접두사 사이의 혼동
(4') 고-멸압-이
(4") 고#멸압-이

위의 '보다' 라는 형태는 조사도 될 수 있고 부사도 될 수 있다. 이러한 품사의 차이는 띄어쓰기를 통해서 드러난다. 또한 '고' 라는 형태의 관형사나 접두사의 차이도 띄어쓰기로 구분되는데 위의 (3")와 (4")는 이러한 품사 표현을 잘못한 예들에 해당한다.

[3] - : 형태소 경계, # : 단어 경계 (어절(語節) 경계)

2.3 발음과 발음 인식에 따른 혼동

다음은 사용자들의 음성학적 혼동과 관련된 예들이다.

(5) 모음 /에/와 /애/ 사이의 혼동
(5') 불에 손을 데고
(5") 차가운 *벽에 손을 데고
(6) 이중모음 /웨/와 /왜/ 사이의 혼동
(6') 웬만하면
(6") *왠만하면
(7) 평음과 경음 사이의 혼동
(7') 쌀#수확; *살#수확
(7") 자장면; *짜장면
(8) 구개음화와 중화현상에 의한 혼동
(8') 굳-이 → *구지
(8") 조예-가 깊다 → *조예가 깁다

현대의 많은 한국어 화자들은 모음 /에/ 와 /애/ 사이의 구분을 잘못한다. 이러한 음성학적 속성이 철자 오류에 그대로 반영되는데 (5')와 같은 경우는 이러한 오류로 인해서 '차가운 벽에 손을 덴다' 는 의미적 모순을 발생시키게 된다. 이와 유사한 혼동이 이중 모음 /에/와 /애/ 사이의 혼동이다. 또한 (7')은 주로 경상도 화자 들에게서 발견된는 혼동 현상이다. 사실상 (7')과 같은 발음 혼동 상의 이유가 철자 인식의 혼동에 까지 영향을 주는 경우는 그리 많지 않다. 그러나 경음에 해당하는 철자를 얻기 위해서는 평음에 해당하는 철자를 얻을 때보다 자판 조작이 어렵다는 사실은 이런 오류 발생률을 높인다. 이에 반해서 (7")의 /ㅈ/ 과 /ㅉ/ 사이의 발음 혼동 현상은 현대의 국어 화자들 사이에서 많이 발견되는 현상으로써 그대로 철자 오류에 반영되어 의미적으로 받아 들일 수 없는 문장을 만든다. 특히 평음과 경음 사이의 혼동 현상은 많은 부분 한국어에 존재하는 경음화(hardening) 현상에 기인한다, 한편, 각각 별개의 형태소로 존재했던 형태들이 합쳐졌을 때 동화 작용 구개음화 등의 음운현상의 영향을 입게 되는 현상은 철자법을 어기는 형태로 드러나게 된다. (8')은 바로 동사 어간 굳~ 'to harden' 에 부사화 접미사 '-이' 가 붙어서 만들어진 어절인데, 출발 형태를 그대로 유지하고 있는 철자법보다는 음성적 변화에 따라 철자법 혼동을 일으킨 경우에 해당하며, (8")은 한국어의 저해음(obstruent)의 내파음(implosive)으로의 중화현상에 기인한 혼동을 보여준다.

2.4 어휘 결합 혼동

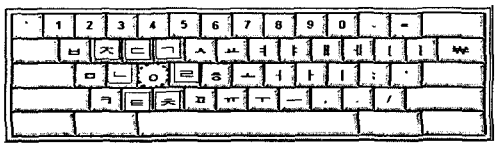
위에서 나열한 여러 원인들 외에도 임의의 두 어휘들이 한 문장에서 같이 나타날 수 있는지에 대한 여부도 문장의 문법성에 영향을 준다.

(9) 위기를#면치#못하고
(9') *위기를#면치#하고
(9") *위기를#금치#못하고

'면치', '금치'는 각각 '면하-지', '금하-지'가 축약되어 어휘화한 형태이다. 여기서 '-지'는 뒤의 부정 의미를 가진 동사연결을 위한 연결 어미이다. 따라서 (9')와 같이 부정의 의미가 아닌 동사가 뒤에 따라 온다면 그 문장은 비문법적인 것이 된다. 또한 '위험'은 누군가가 금하는 것이 아니고 누군가가 면해야 하는 것이다. 따라서 '위험'이란 명사와 '금하지'라는 형태에서 파생된 형태인 '금치'는 한 문장 내에서 양립할 수 없기 때문에 (9")는 비문법적인 것이 된다.

2.5 입력 오류

사용자들은 문서 작성시 타이프 오류를 자주 범한다. 주로 이웃한 자판들 사이에 타이프 오류가 생길 확률이 많다. 예를 들어 '한강'은 '항강'으로 잘못 쓰여질 확률이 높은데, 이때 'ㅇ'과 인접한 자판으로는 'ㄴ', 'ㄹ', 'ㄷ', 'ㅊ', 'ㅈ', 'ㄱ', 'ㅌ'이 있다.



이러한 이웃한 자판들 사이에서도 통계적으로 문제의 자판과 수평 이웃이 수직 이웃 보다 입력 오류일 확률이 높고 수평에서도 왼쪽 이웃이 오른쪽 이웃보다 입력 오류일 높으며 수직 이웃 간에는 해당 자판의 위 자판이 아래 자판보다 입력 오류일 확률이 높다.

2.6 맞춤법 개정에 따른 혼동

표준어 사용 오류나, 순화용어 사용 오류 같은 문체 오류나 관용어 표현 오류, 외래어 표기 오류와 같은 여러 규범 문법관련 오류와 더불어 특기할 사항은 한글 맞춤법 개정안(1989.3) 시행 이후, 종래의 맞춤법에서 표준어로 규정되어 있던 몇몇 형태들이 개정됨에 따라서 개정안 시행 이전에 한국어 철자법을 교육받고 수년에 걸쳐 사용한 사람들에게 고유한 오류가 발견된다.

(9) *있읍니다 → 있습니다
(10) *더우기 → 더욱이

이런 유형의 맞춤법 혼동과 반대되는 현상으로 과잉정정(hypercorrection) 현상이 나타난다. 맞춤법 개정안에 따르려는 노력으로 인하여 (9)에서와 같이 동사의 어간의 마지막 음절이 '쓰'으로 끝나거나 과거 시제 어미 '쓰' 뒤에 종결어미가 따라오는 경우는 '읍니다' 형태를 붙이던 종래의 규칙을 다른 형태의 동사 어간과 통일하여 '습니다'로 통일하여 사용하는 규칙을 지키는데 너무 신경 쓴 나머지 명사화 접미사 '음'을 붙여 파생시킨 형태에도 문제의 개정 규정을 과잉 적용하는 현상이 발견된다.

(11) *있슴 → 있음

3 비대칭 관계에 기반한 부분 문장 분석

이 절에서는 한국어 문장을 분석하는 동안 최선의 결과를 얻기 위해 현 멎춤법 및 문법 검사기(2.0)에 도입된 주요 방법론들을 소개한다. 현 한국어 문법 검사기는, 현재 자연어 이해를 토대로 한 시스템 구축이 어렵다는 점과 완전 문장 분석은 시간이 많이 든다는 문제점을 해결하기 위하여, 부분 문장 분석을 도입하고 있다. 이 분석은 지배소와 피지배소 간의 비대칭 관계를 바탕으로 한 시스템이다. 이때 지배소는 분석 출발점과 분석 방향과 분석 범위를 정의해 준다.

3.1 분석 출발점과 분석 방향

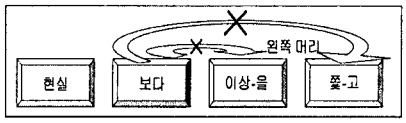
현 부분 문장 분석은 주어진 문장의 특정 부분에서 시작된다. 바로 잠재적 지배소[4]가 분석의 출발점과 분석 방향을 제공해 주게 되는데 이러한 지배소는 의존 문법에서 말하는 지배소 개념과는 차이가 난다. 현 시스템의 지배소는 시스템이 잠재적으로 오류 어절이나 중의적인 어절로 지적할 가능성이 높은 어절이나 구의 패턴을 말한다. 따라서 부분 문장 분석은 시스템에 의해 선택된 지배소로부터 그 지배소의 피지배소까지 이루어진다. 현 시스템의 피지배소는 지배소와 연어관계에 있을 수 있는 어절이나 구 패턴을 일컬을 뿐만 아니라 지배소와 연어관계에 있을 수 없는 어절이나 패턴도 포함한다. 만약 시스템이 연어관계의 어절을 찾는데 지나치게 많은 시간을 소요하고 비연어관계에 있는 피지배소를 찾는 것이 오히려 시간 절약에 도움이 된다면 후자를 선호할 것이다. 이러한 지배소들은 말뭉치와 경험적인 사실들을 토대로 수집되고 정리된 것이다. 현 시스템은 크게 다음과 같은 네 가지 문법 검사 방향을 설정하고 있다.

3.1.1 오른쪽 방향 분석: 한국어 보편 문법의 틀 안에서는 일반적으로 동사가 부사나 보어를 선택하

[4] 잠재적 지배소에 대한 보다 더 자세한 설명을 위해서는 [9]를 볼 것.

는 것으로 주장되고 있다. 본 시스템은 오히려 그 수가 작은 부사나 보어를 시스템의 속도와 성능 향상을 위해서 지배소로서 선호하고 있다. 다음은 위에서 예로 든 3" 의 부분 문장 분석을 보여주는 것이다.

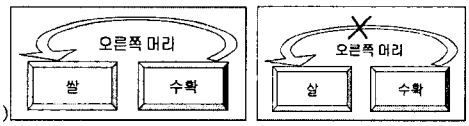
(3")



'보다' 는 지배소로서 부분 문장 분석을 실행시키며, 부사로서 다른 부사나 형용 동사를 지배할 수 있는데 3" 에는 이러한 형태·통사론적 정보를 충족해줄 만한 임의의 피지배소가 없다.

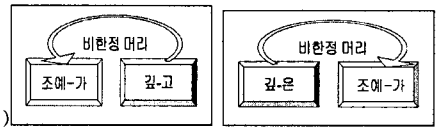
3.1.2 왼쪽 방향 분석: 한국어 복합명사 대부분은 오른쪽에 지배소가 있다. 이러한 언어학적인 특성이 우리 시스템의 분석 방향에도 그대로 반영된다. 예를 들어 (7')의 '살 수확' 이란 형태를 포함하는 문장은 용인될 수 없는 문장으로 지배소인 '수확' 이 임의의 피지배소인 '살' 을 지배하지 못한다.

(7')



3.1.3 양방향 분석: 주어진 지배소의 피지배소 위치가 한정 되지 않은 경우이다. 동사의 보어는 왼쪽이나 오른쪽에 위치할 수 있다. 만약 동사가 관형형 어미와 활용되었다면 분석 방향은 오른쪽이 되며, 그 외 경우에는 왼쪽 방향이 된다. 예를 들어 8" 을 분석하면 지배소로 '깊-고' 가 선택되어 왼쪽 방향으로 분석이 이루어져 피지배소 '조예' 를 찾게 된다. 만약 지배소인 문제의 동사가 관형형 어미로 활용된 경우라면 피지배소가 오른쪽에서 나타나게 되는데 오른쪽 방향으로 분석이 이루어지게 된다.

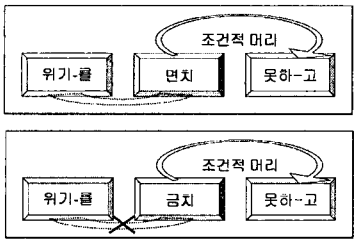
(8")



3.1.4 조건적 지배 관계 분석: 한 통사적 개체가 임의의 피지배소를 지배할 때 그 피지배소가 위치하는 쪽과 반대 방향에 위치하는 다른 요소의 의미론적, 형태·통사론적 조건이 충족되면 지배할 수 있는 지배관계이다. 9를 분석할 때 '면치' 는 지배소로서 오른쪽 방향으로 부분 문장 분석을 실행시킨다. 이때 분석기는 '못하-고' 를 피지배소로 선택한다. (9")를 분석할 때 '금치' 도 지배소로서 '못하-고' 를 피지배소로 선택하기는 마찬가지다. 이 둘은 연어 관계에 있으나 조건적인 지배를 한다. 즉 '면치' 같은 경우는 그 다른 쪽에 '위기-를' 이라는 어절과 연어관계를 이룬다는 조건하에서만 그의 피지배소 '못하-고' 를 지배할 수 있다. 이에 반해서 (9")의 '금치' 는 이러한 조건을 충족해줄 만한 피지배소를 찾지 못한다.

(9)

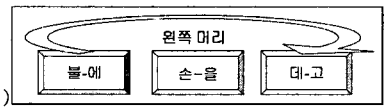
(9")



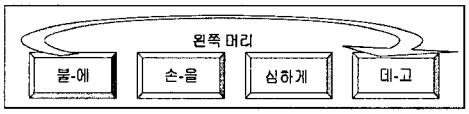
3.2 최대 분석 범위

최대 분석 범위와 문법 검사기에 의해 무시될 수 있는 항목들을 설정함으로써 시스템의 속도를 효과적으로 높일 수 있다. 3.1에서 살펴본 방법으로 지배소가 분석 방향을 제시한다면 분석범위는 지배소의 문법적 위상과 그 지배소가 피지배소 사이의 관계에 의해서 정의된다. 왼쪽에서 오른쪽으로 부분 문장 분석을 하는 (5')를 예로 들어 보자.

(5')



(5')는 다음과 같이 확장될 수 있다:



분석기는 '불-에' 를 지배소로 선택하고 이 어절이 연어관계가 비연어관계에 있을 수 있는 어절이 있는지 검사한다. 그러나 임의의 피지배소를 찾을 때까지 만날 수 있는 부사나 명사는 무시한다.

지배 방향	지배소	피지배소	무시 가능 항목	최대 분석 범위
오른쪽	명사(-조사) 부사	동사-어미	부사 명사	3어절

4 오류 교정 규칙과 처리

이 절에서 살펴볼 언어학적 특성들에 대한 관찰과 이해는 다양한 오류를 다룰 수 있는 분석기의 경험적 단서와 규칙을 구축하는 데 도움을 준다.

4.1 한국어 특성에 기반한 오류 교정 규칙

4.1.1 어절의 순서: 기본적으로 한국어는 주어 + 목적어 + 동사의 순서를 가지고 있는 언어 중의 하나이지만 비교적 자유로운 어순을 가지고 있다. 다음의 통사적 위치는 다소 고정적이다:

- 종결 어미로 끝난 동사는 문장의 제일 마지막에 위치한다.
- 관형어는 수식하는 단어[5] 앞에 위치한다.
- 마침표는 종결 어미로 끝난 동사 뒤에 위치한다.

4.1.2 어절 경계 모형(pattern): 한국어는 일련의 한정된 수의 어미, 조사, 의존명사 등이 있는데, 특별한 형태적 특성을 가지고 있다. 이러한 특별한 형태를 가진 단위들이 만들어 내는 어절간 경계(word bound)의 형태론적 양상은 문서 교정을 위한 단서를 제공하여 준다. 예를 들어 관형절 어미 '-어진', '-던' 등과 연결 어미 '-으니까', '-려고' 등이 다른 통사적 단위들과 더불어 이루는 구 경계 형태는 띄어 쓰기 오류를 위한 다음과 같은 단서를 제공한다:

- 동사 어간-던 # 것
- 동사 어간-려고 # 모든 어절

4.1.3 어미에 반영된 화용론적인 표지: 존대어법의 적절한 사용은 한국어 문장의 문법성을 결정하는 요인이다. 예를 들어, 만약 '교수님' 과 같이 '교수' 라는 명사에 '-님' 과 같은 높임 접미사가 붙고 주격 조사가 붙는다면 그 형태가 속한 문장의 제일 마지막에 오는 주동사는 '-시-' 와 같은 접요사가 붙어야 한다.

- 명사-높임 접미사-주격조사 # (…어절$_i$…) # 동사어간-(시/으시)-어미

4.1.4 음절 특성: 한국어에서 자음군으로 끝나는 대부분의 어휘들은 동사 어간이다. 따라서 띄어쓰기 단위인 어절이 자음군으로 끝나는 경우는 몇 개 없다. 동사가 항상 어미로 활용이 되어야 한다면 자음군이 동사 어절의 마지막에 위치하는 경우가 없다는 말이므로 명사와 같이 어미 접속 없이 별개의 어절로도 문장에서 나타날 수 있는 경우만 관찰 대상에 넣을 수 있다. 대략 '앎', '삶', '값', '닭', '흙' 과 같은 명사들이 여기에 속한다. 이러한 음절 속성을 띄어 쓰기에 사용할 수 있다. 위의 음절을 제외하고 만약 한 음절이 자음군으로 끝난다면 그 다음 띄어쓰기는 불가능하다라는 단서를 잡을 수 있다.[6]

- [# (C)VCC$...] (C)VCC ∉ {앎, 삶, 값, 닭, 흙}

위에서 살펴본 언어학적인 특성 이외의 형태·통사론적 조건들 중에서 동사의 하위범주화를 이용할 수 있다:

- 타동사는 대격조사가 붙은 명사를 뒤따를 수 없다.
- 계사는 아무런 격조사가 붙지 않은 명사에 뒤따라 온다.

4.2 오류 교정 규칙 처리

위에서 살펴본 한국어 특성에 기반하여 만들어진 교정 규칙은 지식베이스 속에 포함된다. 교정 규칙 중 연어 규칙은 하나의 잠재적 지배소와 하나 혹은 그 이상의 피지배소로 이루어진다. 지식베이스 내에서 지배소의 유일성이 보장되어야 한다면 피지배소는 여러 교정 규칙에 포함될 수 있다.

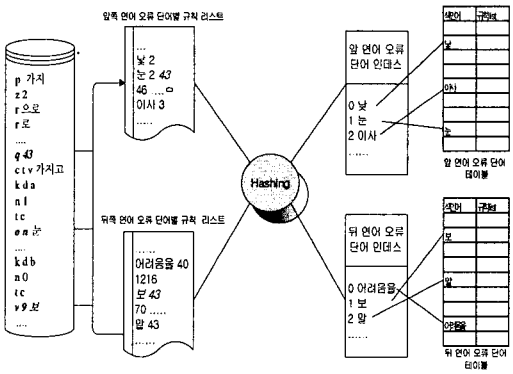
```
규칙(번호) =
  {지배소, 방향, 무시정보, 피지배소의 언어
   학적 조건, 선택}
```

본 시스템은 약 6,000개의 잠재적 지배소나 패턴을 가지고 있어서 이들이 부분 문장 분석을 실행시킨다. 또한 약 5000개의 교정 규칙이 있는데 그 중 2300개 정도의 규칙은 의미·문체적 중의성 제거를 위한 것이다. 규칙들의 수는 분석 속도에 지대한 영향을 줄만큼 많지 않지만 본 시스템은 해싱 함수를 사용하여 이들 규칙을 처리하고 있는데 분석 속도 향상을 도모하고 앞으로의 지식베이스의 확장에 대비하기 위한 것이다. 다음은 지식베이스에서 연어 오류 단어와 해당하는 규칙을 추출하여 의미·문체 사전의 해싱 테이블(Hashing Table) 생성 과정을 보여준다.[6]

---

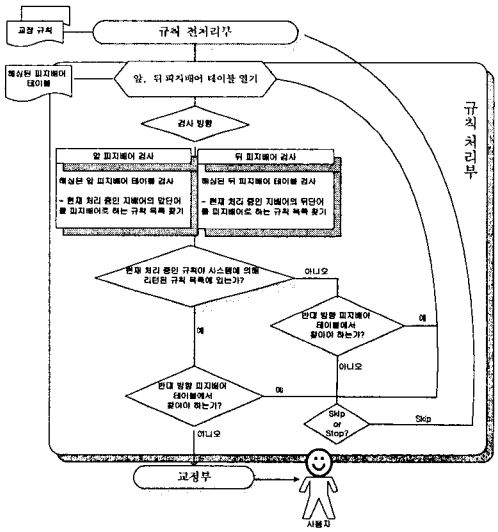
[5] 이 논문에서 사용된 '단어' 라는 용어는 형태·통사론적인 특성에 근거한 정의가 아니라 문장 내에서 띄어쓰기에 의해 구분되는 것에 불과한 것으로 '어절' 과 같은 의미로 사용되었다.

[6] $: 음절경계, C : 자음, V :모음



[그림 1] 해싱 테이블(Hashing Table) 생성 과정

각 규칙은 번호가 부여되어 있다. 잠재적 지배소를 기준으로 앞 또는 뒤로 연어 관계에 있는 피지배소를 검색해 나가므로 잠재적 지배소와 연어 관계 및 연어 오류 관계에 있는 단어별 규칙 리스트는 앞쪽 연어 및 연어 오류 단어별 규칙 리스트와 뒤쪽 연어 및 연어 오류 단어별 규칙 리스트로 분류되어야 한다. 이렇게 분류된 문서는 해싱 기법(Hashing Tool)을 이용해서 연어 및 연어 오류 단어 인덱스와 연어 및 연어 오류 단어 테이블을 생성한다. 연어 오류 단어 인덱스는 연어 오류 단어 테이블의 주소(home address)를 찾는 데 사용된다. 다음은 규칙의 피지배소 즉 연어 및 연어 오류 단어를 검사하는 루틴이다.



[그림 2] 한국어 문법 검사기(2.0) 규칙 처리부

규칙 전처리부에서는 규칙들을 버퍼에 담고 입력된 검사단어 즉 지배소의 조건과 규칙에서 설정된 검사단어의 조건이 같은지 비교해서 같을 때만 규칙을 적용한다.

규칙 처리부에서는 부분 문장 분석을 해서 규칙에 의해 설정된 피지배소 즉 연어 및 연어 오류 관계에 있는 단어를 해싱 기법을 이용하여 검색한다. 문제의 피지배소가 발견되면 교정부에서 대치어와 도움말을 제공한다. *'조건적 지배관계 분석'* 의 경우는 지배소에 따라서 왼쪽 테이블을 먼저 찾을지 오른쪽 테이블을 먼저 찾을지가 정해지는데 현재 검사중인 규칙이 시스템에 의해 반환된 규칙 리스트 속에 있으면 이미 찾은 테이블의 반대 방향 피지배소 테이블에서 찾기를 한다. 반대로 현재 검사중인 규칙이 시스템에 의해 반환된 규칙 리스트 속에 없으면 다음 검사 단어로 옮겨가거나, 멈춘다. 한편, *'양방향 분석'* 의 경우는 검사 중인 규칙이 시스템에 의해 반환된 규칙 리스트 속에 있으면 교정부로 넘어가며 검사중인 규칙이 시스템에 의해 반환된 규칙 리스트 속에 없으면 오른쪽 피지배소 테이블에서 찾기를 한다.

5 결론

이 논문에서 두 임의의 요소들 사이의 비대칭 관계에 기반하여 구축된 문법 검사기를 소개했다. 또한, 오류가 생겨난 언어학적 요인에 대한 이해를 바탕으로 한국어 문장 분석에 적합한 지식베이스를 구축할 수 있었다. 현 시스템의 지식베이스는 주어진 지배소가 그의 피지배소와 연어관계를 유지 할 수 있는지를 검사한다. 즉, 문법 검사기는 오류 가능성이 높은 단어를 지배소로 선정하여 다른 임의의 피지배소와의 지배관계를 설정한다. 잠재적 지배소의 개념을 도입함으로써 현 시스템은 부분 문장 분석의 출발점과, 분석 방향, 그리고 분석 범위를 정의할 수 있었는데 이 모든 것이 시스템 성능 향상에 영향을 주는 것이다. 최근에 본 연구실에서 실시한 테스트에 의하면 현 시스템은 97%의 정확률을 보여준다. 또한 2천만 개의 복합명사를 테스트한 결과 0.1%의 오교정률만을 얻을 수 있었다.

잠재적 지배 관계 개념과 부분 문장 분석 기법을 이용하여 현 시스템이 높은 성능을 얻을 수 있었다고 해도, 무한한 언어사용의 증대에 대한 지속적인 관찰이 필요하며 이를 기반으로 지식베이스를 계속 확장하고 향상시켜 나가야 한다. 또한 사용자의 요구나 기대도 고려해야 하는데, 사용자들의 입장에서 보면 높은 교정률이 곧바로 시스템 성능과 직결되는 것이 아니다. 예를 들어 현 시스템은 비록 옳은 단어가 틀린 것이라고 시스템에 의해 판정된다 하더라도 최대한 많은 오류 단어를 찾는 것에

중점을 둔다. 그러나 사용자들은 자신들의 옳은 단어가 틀리다고 판정된다든지, 아니면 시스템에 의해 틀린 대치어를 제시 받았을 때 더욱 더 민감해진다. 따라서 사용자 입장에서의 문법 검사기의 최적화도 앞으로 계속 연구 되어야 할 과제이다.

6. 참고 문헌

[1]. 강승식. 2002. "한국어 형태소 분석과 정보 검색", 홍릉과학 출판사.

[2]. 김수남. 2000. "운지 거리와 빈도를 이용한 음소 대치의 성능 향상 및 속도 개선", 부산대학교 전자계산학과 석사 청구 학위 논문.

[3]. 김현진. 권혁철. 1998. "어절간 연관 관계를 이용한 한국어 문법 검사기", 정보과학회 논문지, 25-2.

[4]. 소길자, 권혁철. 2001. "어휘적 중의성 제거 규칙과 부분 문장 분석을 이용한 한국어 문법 검사기", 정보 과학회 논문지, 28-3.

[5]. 심철민, 권혁철. 1996. "연어 정보에 기반한 한국어 철자 검사기와 교정기의 구현", 정보과학회 논문지, 23-7.

[6]. 권혁철. 2001. "한중 자동번역을 위한 한글 전처리에 관한 연구", 한국전자통신원 중간 연구 보고서.

[7]. Chae Young-Suk, Kwon Hyuk-Chul. 1991. "A Dictionary-based Morphological Analysis", Proc. of NLPRS ' 91, 141-147.

[8]. Comrie, B. 1989. "Language Universals and Linguistic Typology", Blackwell.

[9]. Kang Mi-Young, Park Su-Ho, Yoon Ae-Sun, Kwon Hyuk-Chul. 2002. "Potential Governing Relationship and a Korean Grammar Checker Using Partial Parsing", Lecture Note in Computer Sience, IEA/AIE . 692-702.

[10]. Kim Hyun-Jin, Park Dong-In, Kwon Hyuk-Chul. 1997. "Implementation of a Korean Grammar Checker Using Collocation of Words and Partial Analysis of a Sentence", Proc. of the IASTED, 208-211.

[11]. Kim Min-Jung, Kwon Hyuk-Chul, Yoon Ae-Sun. 1996. "Rule-based Approach to Korean Morphological Disambiguation Supported by Statistical Method", PACLIC 11, 237-246.

[12]. Kim Su-Nam, Nam Hyun-Sook, Kwon Hyuk-Chul. 1999. "*Correction Methods of Spacing Words for Improving the Korean Spelling and Grammar Checkers*", Proc. 5th Natural Language Processing Pacific Rim Symposium, 415-419.

[13]. Spencer, A. 1991. "Morphological Theory", Blackwell.