형태소 분석기의 어휘적 중의성 해결에 관한 연구

박용욱*

A Study on Lexical Ambiguity Resolution of Korean Morphological Analyzer

Yong-Uk Park*

요 약

한 어절을 중심으로 검사가 이루어지는 맞춤법 검사는 문맥상 어울리지 않는 단어로 인하여 생기는 오류는 찾기 어렵다. 맞춤법 검사기는 현재 어절 단위로 오류 여부를 판단하는 것이기 때문에 어휘적 중의성을 고려하지 않아도 된다. 그러나 문법 검사기는 문장 분석을 해야 하므로 어휘적 중의성을 제거하지 않고는 정확한 검사가 어렵다. 본 논문에서는 어휘적 중의성을 해결하기 위하여 몇 가지 규칙을 만들고 이를 활용하여 문장에 존재하는 어휘적 중의성을 해결할 수 있는 방법을 보인다. 또한 실험을 통하여 그 결과를 분석하였다.

ABSTRACT

It is not easy to find out syntactic error in a spelling checker systems of Korean, because the spelling checker is generally to correct each phrase and it cannot check the errors of contextual ill-matched words. Spelling checker system tests errors based on a words. Disambiguation of lexical ambiguities is important in natural language processing. Its outputs is used in syntactic analysis. For accurate analysis of a sentence, syntactic analysis system must find out the ambiguity of morphemes in a word. In this paper, we suggest several rules to resolve the ambiguities of morphemes in a word. Using these methods, we can reduce many lexical ambiguities in Korean.

키워드

Lexical Analysis, Lexical Ambiguity, Disambiguation of Lexical Analysis
어휘 분석, 어휘적 중의성, 어휘 분석의 중의성 해소

Ⅰ. 서 론

맞춤법 검사 또는 교정기는 일반적으로 한 어절을 대상으로 검사를 하므로 문맥상 어울리지 않는 단어 때문에 발생하는 오류는 찾기 어렵다. 맞춤법 검사기는 현재 어절의 오류 여부를 판단하는 시스템이기 때문에 어휘적 중의성을 고려하지 않아도 된다. 어휘적 중의성이란 하나의 어휘가 두 가지 이상의 의미로 해석될 수 있는 것을 말한다[1,2]. 예를들어 어절 “들어”는 “듣다(동사)+어(어미)”와 “들다(동사)+어(어미)” 두 가지로 해석가능하다. 문법 검사기는 문장 분석을 해야 하므로 이러한 어휘적 중의성을 해결해야만 정확한 검사가 가능하다.

한국어 문장 분석은 언어 지식을 이용하여 자연 언어 문장을 분석하여 여러 개의 후보를 생성하고 생성된 후보들에서 중의성을 문맥과 상황을 이용하여 제

* 울산과학대학교 컴퓨터정보학부
접수일자 : 2012. 06. 22 심사(수정)일자 : 2012. 07. 26 게재확정일자 : 2012. 08. 09

거해 나가는 과정이다. 기호적 방법에 의한 자연 언어 처리는 사전, 문법규칙, 단어 간 호응관계 등을 사용하며, 통계적 방법은 언어의 통계적 특성을 이용한다[3].

자연 언어 처리 단계는 형태소 분석, 구문 분석, 의미 분석 단계로 나눌 수 있디. 형태소 분석은 구문 분석이나 의미 분석의 전처리 단계이며, 형태소 분석의 정확도나 속도는 상위 처리 단계인 구문 분석이나 의미 분석의 정확도나 속도와 밀접한 관련이 있다[5].

굴절어인 영어는 어절이 대부분 한 실질 형태소에 한두 개의 접사가 붙어서 형성되므로, 사전 탐색에 간단한 접사 처리로 형태소 분석을 할 수 있다. 영어 형태소 분석은 사전과 간단한 변형 규칙만을 이용하므로 형태소 분석기의 구현이 간단하다. 그러므로 영어 형태소 분석은 비교적 쉬운 편이다.

한국어 형태소 분석은 영어 형태소 분석기에 비하면 분석이 어렵기 때문에 속도가 느리다. 그리고 응용 시스템에 따라서는 접미사, 접두사, 보조용언 처리 및 개별 어휘 정보를 활용한 정확한 분석을 요구하기도 하나, 대부분의 한국어 형태소 분석 시스템은 처리 속도와 기술적인 문제 때문에 접미사와 접두사에 대한 정확한 분석을 하지 않고 있다. 하지만 한국어 형태소 분석 시스템을 실용화하고, 다양한 응용 시스템에 활용하기 위해서는 접두사 및 접미사 분석을 포함한 정확한 분석과 빠른 속도 그리고 신뢰도가 높은 중의성 제거가 요구된다[4].

II. 형태소에 존재하는 중의성

형태소는 일정한 음성에 일정한 뜻이 결합되어 있는 말의 가장 작은 단위이다. 즉 최소 어미 단위이다. 형태소 분석이란 주어진 문장을 구성하는 어절들에 들어 있는 최소의미 단위의 형태소들을 분리하여 내고, 분리한 형태소간의 결합 관계를 분석하여 불필요한 분석을 제거하며, 형태소의 원형을 복구하는 과정을 말한다. 이러한 형태소 분석은 구문 분석을 하기 위한 전단계로 실시된다. 형태소 분석의 결과는 구문 분석기의 성능에 큰 영향을 주므로 정확한 형태소 분석은 매우 중요하다[7].

한국어는 형태상 교착어로 분류되며, 계통상으로는 알타이어족에 속한다. 교착어는 의미를 나타내는 실질 형태소에 조사나 어미와 같은 형식 형태소가 붙어서 각 어절의 문장 내에서 문법 기능을 결정한다. 교착어는 부착어나 첨가어라고도 하며, 한국어 외에 일본어, 터키어, 몽골어 등이 이에 속한다. 교착어는 형태소의 결합 구조가 매우 복잡하다[6,9].

자연 언어 중의성은 어휘적 중의성, 품사적 중의성, 구조적 중의성, 의미적 중의성으로 분류된다. 어휘적 중의성은 형태소의 원형을 분리하고 복원하는 과정에서 발생하며, 품사적 중의성은 하나의 형태소가 여러 가지의 품사로 사용되므로 발생한다. 구조적 중의성은 한 문장이 다수의 통사구조로 분석될 때 발생하며, 의미적 중의성은 단어의 의미 해석이 두 가지 이상으로 가능할 때 나타난다. 구조적 중의성과 의미적 중의성은 구문 분석과 의미 분석에서 다루어야 하므로, 형태소 분석 과정에서 발생하는 중의성인 어휘적 중의성과 품사적 중의성에 대해서만 본 논문에서 다루었다.

다음은 어휘적 중의성과 품사적 중의성의 예이다.

① 어휘적 중의성
"철수가 사과를 먹은 줄 안다" 문장에서 (먹은)은 다음과 같이 두가지 분석 가능하다
- 먹은 : 먹다(동사) + 은(관형형어미)
- 먹은 : 먹(명사) + 은(조사)

② 품사적 중의성
"영희가 그네를 탄다" 문장에서 (그네)는 다음과 같은 두가지의 품사로 해석 가능하다.
- 그네 : 명사
- 그네 : 대명사

이와 같은 어휘적 중의성 및 품사적 중의성은 구문 분석이나 의미 분석을 모두 거쳐야만 바르게 해결될 수 있으나, 본 논문에서는 좌우 어절 간의 연관 관계 및 문법관계를 이용하여 상당한 수준에서 이러한 형태소 중의성을 제거할 수 있음을 보인다.

III. 제안된 형태소 중의성 제거 과정

본 논문에서는 어휘적 중의성을 해결하기 위하여

몇 가지 규칙을 만들고 이를 활용하여 문장에 존재하는 어휘적 중의성을 해결 할 수 있는 방법을 설명한다. 두 종류의 중의성 제거 규칙과 통계적 정보에 의한 보완 방법을 사용한다. 또한 문장에 존재하는 오용어는 표준어로 대치한 후 형태소 분석을 통해서 중의성을 제거한다. 규칙 적용은 중의성이 존재하는 어절 및 좌우 어절에 대해 중심적인 역할을 하는 지배소 어절을 결정하고 이를 통해 중의성을 제거한다. 지배소 어절을 결정하는 방법에는 두 가지 방법이 있다. 또한 이를 활용할 수 있는 규칙을 만들어 중의성을 제거한다.

3.1 어휘적 중의성 제거 규칙

어휘적 중의성을 제거하는 규칙에는 크게 두 가지로 분류했다. 첫 번째는 형태소 분석시에 중의성을 발생시키는 빈도가 높은 형태소에 대한 것이고, 두 번째는 형태소 좌우의 연관관계에 대한 규정이다.

먼저 중의성 발생빈도가 높은 형태소에 대한 규정을 정한 (rule1)은 다음과 같다.

(rule1) 중의성 발생빈도가 높은 형태소에 대한 규칙들

(rule1.1) (체언)+가/이/도
~ 있다/없다/아니다(형용사)
(rule1.2) (관형어) + 수(의존명사)
~ 있다/없다(형용사)
(rule1.3) 과/와(접속조사)
앞뒤로 대등한 자격을 갖는 낱말이나 어절이 옴

(rule1.1)은 “있다/없다/아니다” 앞에는 체언+“가/이/도” 조사가 와야하고, 이때 “있다/없다/아니다”는 형용사가 된다는 규칙이다. (rule1.2)는 “수” 앞에는 관형어가 와야 하고, 이때 “수”는 의존명사이고, “있다/없다”는 형용사가 된다는 규칙이다. (rule1.3)은 접속조사 “과(와)”에 대한 규칙으로 “과(와)” 앞뒤에는 같은 자격을 같은 낱말이나 어절이 와야 한다는 규칙이다. (rule1)에 해당하는 규칙은 어휘유형을 계속 조사하여 추가 할 수 있다.

(rule1)은 중의성 빈도가 높은 특정 형태소에 대한 규칙인데 반해 (rule2)는 보다 일반화된 규칙이다. 형태소들의 좌우 연관관계 제약에 대한 일반적인 사항을 규정한 (rule2)의 내용은 다음과 같다.

(rule2) 형태소 좌우 연관관계에 대한 구문 분석 제약 규칙들

(rule2.1)의존명사는 관형어 뒤에 와야 함
(rule2.2)수사 의존명사는 수관형사 뒤에 옴
(rule2.3)관형어 뒤에 관형어나 수식을 받는 명사가 와야 함
(rule2.4)보조용언 앞에는 본용언 와야 함
(rule2.5) 목적어는 자동사나 형용사 앞에 올 수 없음
(rule2.6) 감탄사와 접속부사는 문장의 처음에 와야 함
(rule2.7) 관형어 뒤에는 부사어가 올 수 없음
(rule2.8) 수관형사 뒤에는 수사 의존명사와 명사, 관형어가 올 수 있음
(rule2.9) 수사 어절은 순서가 있음

이러한 규칙을 사용하여 다음의 예문을 처리하는 과정을 살펴보겠다.

예문a) 학교에 갈 수가 없다

이 예문에서 “갈 수가”에 대해 분석하면 아래와 같다.

=> 갈 : 갈(명사)
갈다(동사)+ㄹ(관형형어미)
수가 : 수(명사)+가(조사)
수(의존명사)+가(조사)

(rule1.2)에 의하여 의존명사 “수”는 앞에 관형어가 와야 하기에 앞의 “갈다(동사)+ㄹ(관형형어미)와 결합가능하다. 따라서 어절 ”갈“은 갈다(동사)+ㄹ(관형형어미)로 해석되고, 어절 ”수“는 수(의존명사)+가(조사)로 해석된다. 또한 (rule2.1)에 의해서도 동일한 결과를 얻을 수 있다.

예문b) 엄마가 보고 싶다

이 예문에서 “보고 싶다”에 대해 분석하면 아래와 같다.

=> 보고 : 보고(명사)
보다(동사)+고(연결어미)
싶다 : 싶다(보조용언)+다(어미)

이 예문은 (rule2.4)을 적용하여 중의성을 제거할 수 있다. (rule2.4)에 의하면 보조용언 앞에는 본용언이 와야 한다. 따라서 연결어미 “고”가 붙은 본용언인 “보다(동사)”와 연결되는 것으로 분석된다. 따라서 어절 “보고”에 존재하는 중의성은 “보다(동사)+고(연결어미)”로 분석된다.

3.2 지배소 어절 지정 및 규칙을 사용하여 중의성 제거

지배소 어절을 지정하여 중의성을 제거하는 방법에는 다음과 같은 두 가지가 있다.

① 중의성을 가진 어절을 지정하여 중의성 제거
② 중의성을 가지지 않은 좌우 어절을 지정하여 중의성 제거

①의 방법은 중의성을 가진 어절을 중심으로 좌우 어절을 살펴보아 중의성을 해결하는 방법이다. “성공할 수 있다면” 이라는 문장에서 살펴보겠다. 이 문장에서 어절 “수”는 명사와 의존명사 두 가지로 분석 가능하다. 이때 어절 “수”는 중의성 제거의 지배소 어절이 되고 좌우 어절인 “성공할”과 “있다면”의 분석 결과를 이용하여 “수”어절의 중의성을 해결한다. “수”가 의존명사가 되려면, 규칙에 의해 왼쪽에 관형형 어미가 와야 하며 오른쪽에는 “있다/없다”가 와야한다. 따라서 이 문장에서 “수”어절은 의존명사로 해석되게 된다.

②의 방법은 중의성이 없는 어절이 중심이 되어 좌우에 존재하는 중의성 어절을 해결하는 방법이다. “아이가 먹지 않고” 라는 문장으로 살펴보겠다. 이 문장에서 먹지는 두 가지로 분석 가능하다. 하나는 명사(먹지)로 분석되고 다른 하나는 먹다(동사)+지(어미)로 분석되는 것이 가능하다. 이때 중의성이 없는 어절 “않고”를 지배소 어절로 정하여 앞에 있는 어절 “먹지”의 중의성을 제거한다. 어절 “않고”는 “않다”(보조용언)+고(어미)로 해석된다. 지배소 어절인 “않다”는 보조용언으로서 왼쪽에 어미 “-지”로 끝나는 용언과만 결합한다. 따라서 “먹지(명사)”와 “먹다(동사)+지(어미)”로 중의성을 갖는 어절은 지배소 어절 “않다”에 의해 “먹다(동사)+지(어미)”를 올바른 것으로 분석한다.

또한 본 논문에서는 형태소 분석을 하는 과정에서 발생할 수 있는 오용어에 대하여 처리를 한다. 형태소 분석과정에서 오용어가 나타나면 이것을 표준어로 대치하고 대치된 표준어에 대하여 다시 형태소 분석을 하여 중의성을 제거하는 방법을 적용하였다. 오용어 처리 과정은 먼저 오용어를 발견하고, 발견된 오용어에 대하여 오용어 데이터베이스를 이용하여 표준어를 찾아 교정하는 순으로 진행한다.

다음은 오용어가 있는 문장에 대하여 오용어를 표준어로 대치하고, 오용어가 제거된 문장에 대하여 중의성을 제거하는 과정을 보여주는 예이다.

예문c) 먹지 않타.

먼저 “않타”가 “않다”에 대한 오용어이므로 이를 인식하여 오용어처리 데이터베이스를 활용하여 “않다”로 대치하게 된다. 그리고 이제 오용어가 제거된 결과의 문장인 “먹지 않다”에 대하여 분석한다. 그 과정은 다음과 같다.

=> 먹지 : 먹지(명사)
먹다(동사))+지(어미)
않다 : 않다(보조용언)+다(어미)

결과적으로 “않다”는 한가지로만 분석됨으로 문제가 없으며, “먹지”는 두 가지로 분석되어 중의성을 제거해야 한다. 형태소의 좌우 연관관계에 대한 구문분석 제약 규칙 중에 보조용언 앞에는 본용언 와야 한다는 규칙인 (rule2.4)가 있다. 따라서 이 (rule2.4)를 적용하면, “먹지”에 대한 두 가지 분석결과인 “먹지(명사)”와 “먹다(동사)+지(어미)” 중에서 용언이 될 수 있는 “먹다(동사)+지(어미)”로 해석된 것을 취할 수 있어 중의성을 해결한다.

Ⅳ. 실험 및 결론

이 장에서는 본 논문에서 제안한 중의성 제거 규칙을 사용하여 성능을 평가한다. 평가에 사용한 데이터는 중학교 교과서이다. 평가는 다음의 정확도 식을 사용한다.

$$\text{정확도} = \frac{\text{올바르게 분석된 어절 수}}{\text{전체 어절 수}}$$

표 1. 실험결과
Table 1. Experimental Results

실험대상	어절수	중의성 이있는 어절	처리된 어절	정확률
중학교 교과서	2487	1342	1218	1218 /2487 (90.7%)

중학교 교과서에서 얻은 2487 어절 중에서 1342개의 어절이 중의성이 있는 어절이고, 실험결과 이중 1218 어절이 규칙 적용에 의해 바르게 분석되어 정확도는 90.7%이다.

더 좋은 정확도를 얻기 위해서 중의성 제거 규칙의 수를 계속해서 확장해야 하며, 또한 규칙을 보다 정확하게 적용할 수 있도록 규칙을 개선해야 하는 작업이 진행되어야 한다.

본 논문은 자연 언어 처리의 제일 선행단계라 할 수 있는 형태소 분석단계에서 발생하는 어휘적 중의성을 해결하는 방법으로 규칙을 적용하여 실험하였다. 형태소 분석에서 중의성이 해결되지 않으면, 이것은 다음단계인 구문 분석의 복잡도 증가로 그대로 전달됨으로 형태소 분석단계에서 해결할 수 있는 어휘적 중의성을 가능한 많이 해결해 주는 것이 자연 언어 처리 시스템 개발에 매우 중요하다[2]. 중의성은 구문 분석이나 의미 분석을 모두 거쳐야만 완벽하게 해결될 수 있지만 본 논문에서는 좌우 어절 간의 연관 관계를 적용한 규칙을 만들어 중의성을 해결하여 상당 수준 해결 할 수 있음을 보였다.

향후 중의성을 제거할 수 있는 규칙을 보다 많이 추가해야 하고, 규칙을 정교하게 처리할 수 있는 알고리즘을 개발하는 과정이 필요하다.

참고 문헌

[1] 남기심, 고영근, “표준 국어 문법론”, ㈜탑출판사, 2009.

[2] 강승식, 김영택, “사전 정보에 기반한 효율적인 한국어 형태소 분석기의 설계 및 구현”, 한국정보과학회 봄학술발표논문집, 18권, 1호, pp. 529-532, 1991.

[3] 김영택, 권혁철, 옥철영, 서영훈, 이호석, 이근배, 윤덕호, 문유진, 강승식, 이하규, 심광섭,윤성희, 서병략, 이재원, 장병탁, 양재형, 양승현, 김성동, 김유섭, 이종우, 오장민, 박성배, 장정호, 황규백, 신형주 공저, “자연언어처리”, 생능출판사, 2001.

[4] 조재현, 양황규 “컬러 정보 및 형태학적 특징과 신경망을 이용한 차량번호판 인식”, 한국전자통신학회논문지, 5권, 3호, pp. 304-308, 2010.

[5] 한길, “우리말의 융합 형태소와 형태소 중복현상”, 강원대학교 인문과학연구소, 인문과학연구, 15집 pp. 55-84, 20006.

[6] 심광섭, “MADE : 형태소 분석기 개발환경”, 한국인터넷정보학회, 인터넷정보학회논문지, 8권 , 4호 pp. 159-171, 2007.

[7] 조재현, “자동번호판 이진화를 위한 개선된 퍼지 이진화 방법 ”, 한국전자통신학회논문지, 6권 2호, pp. 231-235, 2011.

[8] 박용욱, “의존문법 기반의 구간 분할법을 황용한 한국어 구문분석기”, 한국해양정보통신학회논문지, 13권, 8호, pp. 1705-1711, 2009.

[9] 김분희, “메뉴와 쇼셜 네트워크 공동구매 정보 도시제공 P2P 시스템”, 한국전자통신학회논문지, 6권, 3호, pp. 445-449, 2011.

저자 소개

박용욱(Yong-Uk Park)

1991년 부산대학교 전자계산학과 (이학석사)
1991년 3월～1997년 2월 전자부품연구원(KETI) 전임연구원
2000년 부산대학교 전자계산학과(박사수료)
1998년 3월～현재 울산과학대학 컴퓨터정보학부 교수