産業科學技術硏究所 論文集 第十一卷 第二號
(J. of Ind. Sci. and Tech. Institute)
Vol. 11, No. 2, 1997. 12.

전처리와 조사/어미의 결합 제약을 이용한 한국어 형태소 해석의 중의성 해소

김 남 철, 정 천 영, 서 영 훈
충북대학교 컴퓨터공학과

The Morphological Ambiguity Resolution for Korean Analysis Using Pre-Processing and Combination Restriction of Josa/Eomi

Nam-Churl Kim, Cheonyoung Jung, Young-Hoon Seo
Chungbuk National University.

Abstract

There are many morphological ambiguities for analysing Korean language, such as part-of-speech ambiguity, stem-type ambiguity, morpheme segmentation ambiguity, morpheme-length ambiguity, irregular stem ambiguity, morpheme-drop ambiguity. If the ambiguities are not resolved in the morphological analysis step, the considerable amount of ambiguities will be generated in the next analysis step, for example, syntactic analysis, semantic analysis.

In Korean, some josas and eomis need specific stems. This characteristic is useful for disambiguation of analysis. In this paper, the combination restriction of josa/eomi is used for disambiguating Korean analysis. The pre-processing of Korean analysis is proposed for analysing properly the word that is mixed with Korean and non-Korean characters or is including the quotation marks, parenthesis marks, or bracket marks.

In many words that have the morphological ambiguity, the ambiguities were resolved, and the average success rate was 92%.

Ⅰ. 서 론

한국어 형태소 분석 결과는 교착어라는 특성 때문에 영어와 같은 굴절어의 분석 결과와 비교할 때 그 유형 및 중의성 유형이 매우 다양하다 [1,2]. 그것은 영어의 형태론적 중의성이 품사 중의성으로 규정되는데 비해, 한국어의 경우는 품사 중의성뿐만 아니라 형태소 분리 위치, 단어의

유형, 형태소의 원형 복원, 탈락 및 축약 현상 등을 처리하는 과정에서 여러 가지 중의성이 발생하기 때문이다[3,4].

그런데 형태소 분석의 결과는 형태소 분석 자체로서 처리가 끝나기보다는 다음 단계의 응용 처리 시스템의 입력으로 제공되기 때문에, 형태소 분석 단계에서 발생된 중의성 문제를 해소하지 않는다면, 이 결과를 입력으로 받는 응용 처리 시스템의 부담은 훨씬 커지게 되며, 분석을 문장 단위로 처리하는 구문 분석이나 의미 분석 단계에서는 각 어절들의 모든 가능한 조합을 검사하여야 하는 경우가 생기므로 중의성 문제는 기하급수적으로 증가된다.

따라서 형태소 분석 단계에서 발생한 중의성은 형태소 분석 단계에서 해소하는 것이 가장 바람직하다. 하지만 형태소 분석은 어절을 기본 처리 단위로 하기 때문에 어절 자체만으로는 중의성 해소가 어려운 경우가 많다.

지금까지 연구된 형태론적 중의성을 해소하기 위한 방법론들은, 태깅된 말뭉치(Tagged corpus)를 구축하기 위한 과정으로부터 연구되었으며, 대부분이 통계적 기법을 이용하였다[5,6,7]. 그러나 통계적 기법에서는 자료 영역에 따른 의존도가 높고, 자료 부족(Data Sparseness)문제뿐만 아니라 미등록어 처리에 따른 문제도 어려운 문제로 남아있다. 또한 태깅된 말뭉치를 구축하기 위한 과정으로부터 연구가 진행되었기 때문에, 형태론적 중의성이 발생하는 원인을 파악하고 유형을 분류하여 중의성을 해소하고자 하는 분석적 접근 방법에 대한 연구는 미흡하였다.

본 연구에서는 형태소 분석시 발생하는 중의성 중 조사/어미와 체언/용언 결합시 발생하는 중의성을 해소하기 위하여 조사/어미의 결합 제약을 이용한다. 또한 선행어절 및 후행어절의 제약을 통한 중의성 해소도 제시하며, 한 쌍으로 나타나는 문장부호인 따옴표 및 묶음표와 한영 혼용어에 대한 올바른 처리를 위한 전처리를 마련한다.

II. 형태소 분석의 전처리

형태소 분석기의 전처리 기능으로서 한 쌍으로 나타나는 문장 부호와 한영 혼용어에 대한 적절한 처리 기능을 부여한다.

1. 따옴표와 묶음표에 대한 처리 방안

따옴표와 묶음표에 대한 처리에 있어서 많은 형태소 분석기가 다른 문장부호의 처리와 동일하게 취급하기 때문에, 이런 문장부호들은 어절이 분리되는 위치 제공의 역할밖에 하지 못한다. 다음은 따옴표와 묶음표가 문장 내에서 사용되는 예이다.

(2-1) "밥 먹었니?"
(2-2) 스스로 "청나라를 따라 죽는다"고 선언 …
(2-3) 미국에서 출판한 《중국철학소사》에서 …
(2-4) 미국학자인 퍼스(Charlotte Furth)는 …
(2-5) 수학적 시간이 선(線)이라는 것을 …
(2-6) 그 문장에서 쓰인 낱말[單語]에 대한 …

문장 (2-1)~(2-6)에서 밑줄 친 부분 중 '고', '에서', '는', '선', '이라는', '에'에 대한 형태소 분석을 수행하면 다음과 같은 중의성을 나타낸다.

(2-2R) 고(체언); 고(관형사); 고(조사)
(2-3R) 에(용언)+서(어미); 에서(조사)
(2-4R) 늘(용언)+ㄴ(어미); 는(조사)
(2-5R) 서(용언)+ㄴ(어미);
설(용언)+ㄴ(어미);
선(체언)
이(용언)+라는(어미);
이(체언)+이(서술격조사)+라는(어미);
이라는(조사)
(2-6R) 에(용언)+어(조사); 에(조사)

위와 같은 분석은 따옴표와 묶음표에 대한 문장부호의 의미를 전혀 고려하지 않은 결과이다. 따라서 자주 발생하는 이러한 문장부호들에 대

한 올바른 처리를 통하여 중의성을 해소할 수 있다.

큰따옴표(" ")는 대화, 인용, 특별 어구 따위를 나타낼 때 사용되는 문장부호이며, 작은따옴표(' ')는 따온 말 가운데 다시 따온 말이 들어 있을 때에 쓰이거나 마음속으로 한 말을 적을 때와 특정한 어구에 대한 강조를 위해 쓰인다.

소괄호(( ))는 원어, 연대, 주석, 설명 등을 넣을 경우에, 중괄호({ })는 여러 단위를 동등하게 묶어서 보일 때, 대괄호([ ])는 묶음표 안의 말이 바깥 말과 음이 다를 때 및 묶음표 안에 또 묶음표가 있을 때에 쓰인다[8].

그러나 이러한 맞춤법 규정이 있더라도 정확하게 꼭 그 경우에만 이런 문장부호들이 사용되는 것은 아니다. (2-3)에서와 같이 책제목이나 글의 제목을 표시 할 때 중괄호나 대괄호를 사용하기도 한다. 따라서 따옴표와 묶음표에 대한 실제적인 사용이 일반 문서에서 어떤 유형으로 사용되는가에 따라 가능한 모든 경우에 올바른 처리를 하는 것이 필요하다.

따옴표와 묶음표가 문장에서 나타나는 유형은 문장부호 좌우로 따라 붙는 어절의 유무에 따라 다음과 같이 나눈다.

(유형1) 따옴표나 묶음표의 앞뒤에 다른 어절이 따라 붙지 않는 경우: (2-1)
(유형2) 따옴표나 묶음표의 뒤에 어절이 따라 붙는 경우: (2-2), (2-3)
(유형3) 따옴표나 묶음표의 양쪽에 어절이 따라 붙는 경우: (2-4), (2-5), (2-6)

여기서 어절이 따라 붙는다는 것은 문장부호와 어절간 공백 없이 나타나는 것을 의미하며, 형태소 분석시 (유형1)의 경우는 별다른 처리가 요구되지 않지만 그 외의 유형에 대해서는 다음과 같은 처리 방안이 필요하다.

따옴표와 묶음표에 대한 처리

(유형2)는 따옴표나 묶음표 어절 전체를 하나의 명사로 취급하고 뒤에 따라 붙는 어절을 명사와 결합될 수 있는 어절로 처리한다.

(유형3)은 따옴표나 묶음표 어절을 한 어절 사이에 삽입된 것으로 간주하여 앞뒤의 어절을 붙여 하나의 어절로 만든 후 처리한다.

각각의 경우 따옴표나 묶음표 어절은 별도의 독립된 어절이나 문장으로 처리한다.

위의 처리 방법에 의해 문장 (2-2)~(2-6)을 형태소 분석하면, 다음과 같이 (2-5)를 제외한 나머지에 대해서는 완벽하게 중의성이 해소되며, (2-5)에 대해서도 올바른 2개의 분석 후보를 제시한다.

(2-2R2) " "(체언) + 고(조사)
(2-3R2) { }(체언) + 에서(조사)
(2-4R2) 퍼스(체언) + 는(조사)
(2-5R2) 선(체언) + 이라는(조사);
선(체언) + 이(서술격조사) + 라는(어미)
(2-6R2) 날말(체언) + 에(조사)

2. 한영 혼용어의 처리

한글 문서에 영어가 나타나는 유형을 살펴보면 다음과 같이 네 가지 경우로 요약된다.

(유형4) 영어로만 구성된 경우: Search, ISO
(유형5) 영어+한글로 구성된 경우:
LAN의, TV는, CDMA이동통신의
(유형6) 한글+영어로 된 경우:
한국IBM, 노트북PC
(유형7) 한글+영어+한글의 경우:
한국MS사는, 윈도즈NT는

이 중 가장 많이 나타나는 유형은 (유형5)의 경우이며, (유형6)과 (유형7)의 경우는 주로 신문이나 기술 서적에서 나타난다. 네 가지 유형 모두 영어는 명사의 기능으로 쓰이거나 또는 복합명사의 단위명사로 쓰인다. 따라서 다음과 같은 처리 기준으로 영어와 한글이 결합된 어절에 대한 처리를 할 수 있다.

한영 혼용어의 처리

한영 혼용어에서 영어는 모두 명사의 기능, 또는 복합명사를 구성하고 있는 단위명사의 기

능으로 분석하고 이에 따른 나머지 어절을 분석한다.

위의 처리 방법에 따라 한영 혼용어를 처리할 경우 각 유형에서 보이는 단어들은 올바르게 처리가 된다.

III. 조사/어미의 결합 제약 강화

국어 사전[9]에 수록된 조사의 수는 현재에는 거의 쓰이지 않는 옛 말을 제외하고 약 140개이다. 조사는 조사끼리 서로 결합할 수가 있으며, 적게는 두 개의 조사결합이, 많게는 '-에서처럼만큼만이라도'와 같이 다섯 개의 조사결합이 가능하다. 두 개 이상이 결합된 조사까지 모두 포함하여 현재 수집된 조사의 수는 약 2,160개가 된다[ETRI 조사, 어미 사전].

또한 사전에 수록된 어미의 수는 495개인데, 조사와 결합 가능한 것을 포함하여 수집된 어미는 약 800개이며, 선어말 어미와 어말 어미가 결합된 형태까지 모두 합하면 3,728개가 된다[ETRI 조사, 어미 사전].

이들 조사와 어미 중 같은 형태를 가지는 형태소 즉, 하나의 형태소가 조사와 어미 두 가지로 쓰이는 형태소 모두 17개로 다음과 같다.

게, 고, 나, 나마, 는, 다, 다가, 든, 든지, 라고, 라도, 며, 아, 요, 은, 을 이

그런데 이 들 형태소와 결합하는 어휘 형태소가 품사 중의성 즉, 어간의 품사가 체언과 용언이 될 수 있는 단어라면 형태소 분석시 중의성을 발생시킨다.

사전에 등록된 어휘 형태소를 조사해 본 결과 동일한 어휘 형태소에 대하여 체언과 용언의 품사가 함께 존재하는 수는 694개이다. 따라서 이들 형태소와 어미-조사 중의성을 갖는 17개의 문법 형태소에 의해 중의성이 발생할 수 있는 전체 어절의 갯수는

(중의성 어간 수) × (중의성 문법 형태소 수) = 694 × 17 = 11,798 이 된다.

위의 조사-어미 중의성을 갖는 문법 형태소 중 가능한 조사-어미의 중의성을 제거하면 원천적으로 존재하는 중의성을 해소할 수가 있다.

또한 어떤 조사나 어미는 체언이나 용언과 결합할 때 매우 제한적으로 사용되는 경우가 많다. 이를 효과적으로 적용하면 형태소 분석시 많은 중의성을 해결할 수 있다.

1. 어절 자체 내에서의 결합 제약

가. 조사 '-게'의 제한적 사용

문법 형태소 '-게'는 어미와 조사로 함께 쓰이지만 조사로서의 '게'는 조사 '-에게'의 준말로 매우 제한적으로 사용된다. 즉 '내/네/제/우리/너희/저희'와 결합될 수 있으며 그 이외의 체언과는 결합하지 않는다. 따라서 '게'는 조사 사전에서 제거하고, 어근이 위의 '/내/네/제/우리/너희/저희'일 때 조사로서의 결합여부만 처리함으로써 불필요한 후보 생성을 방지할 수 있다.

나. 조사 '-고'의 낮은 선호도 반영

어미로서 '고'는 두루 쓰이나, 조사로서 '고'는 '모음으로 끝나는 체언 아래에서 두 가지 이상의 사물을 아울러 말할 때 쓰이는 접속 조사'로 사용 예는 다음과 같다.

(3-1) 서로 어깨를 기대고(어미) 있는 모습이 보인다.
(3-2) 공부고(조사) 뭐고(조사) 다 그만두어라.

그러나 말뭉치 분석 결과 약 45만 어절 중 '고'가 문법 형태소로 쓰인 어절은 14,311어절이었으며 이 중 '고'가 조사로 쓰인 어절은 한 번도 없다. 하지만 어간 부분이 체언-용언 중의성이 발생하여 '고'가 조사로 분석되어 중의성을 나타낸 횟수는 3,260에 해당된다.

따라서 조사 '고'는 일반 조사와 동일시하지 않고, 일반적인 분석 후 결과가 없을 경우에만

검사하는 방법을 사용하며 우선 순위도를 낮추면 문법 형태소 '고'에서 발생하는 중의성을 해소 할 수 있다.

다. 조사 '-다'와 '-다가'의 사용 제약

조사로서의 '-다'는 '-다가'는 모두 '-에다가'의 준말로 사용 예는 다음과 같다.

(3-3) 여기다 묻어라. 책은 저기다 묻어라.
(3-4) 저기다가 책가방은 두어라.

일반적으로 장소를 나타내는 체언에는 모두 결합할 수 있겠으나, 실제 말뭉치를 조사한 결과 29만 어절 중 '-다'와 '-다가'가 조사로 쓰인 어절은 2개이며 '거기다'와 '거기다가'였다.

또한 조사 '-에다'나 '-에다가'의 의미로서 사용될 경우에는 '-에-'가 거의 생략되지 않으며 29번 발생된 어절에서 '-에-'가 생략된 경우는 한 번도 없었다.

그러므로 조사 '-다'와 '-다가'는 조사 사전에서 특수조사로 간주하고 어근이 '여기', '저기', '거기' 인 경우에만 조사로 처리하며, 그 외에는 어미로 처리함으로써 중의성을 해소할 수 있다.

라. 어미 '-이'의 제한적 사용

문법 형태소 '-이'는 조사와 어미가 모두 있다. 조사 '-이'는 주격조사 및 보격조사로 매우 빈번히 사용되는 반면, 어미 '-이'는 모음으로 끝나는 형용사 어간에 붙어서, '하게' 할 자리에 자기의 생각한 바를 말할 때 쓰이는 종결 어미로서 사용되는 경우가 매우 제한되어 있다.

(3-5) 자네 솜씨가 정말 대단하이.

그런데 '-이'가 어미로서의 비중과 조사로서의 비중이 동등하게 등록이 되어 있어서 중의성이 발생되는 문장은 상당히 많다. 문법 형태소 '-이'가 사용된 8878어절 중 571어절에서 중의성이 발생하였다. 따라서 1차적인 처리에서 분석된 결과가 없을 경우에만 '-이'를 어미로 간주하여 분석을 시도하는 방법을 사용하는 것이 중의성 해소를 위한 방법이다. 또한 '-이'가 어미로 쓰이는 경우는 문장의 맨 마지막에 온다는 조건도 적용해야 한다.

마. 어미 '-라', '-어라'와 같은 명령형 어미의 제한적 사용

일반적인 문서에서 명령형 어미가 나타나는 예는 거의 없으며, 구어체 문장이 나타나는 문서에서만 발견되는 경우가 많으며, 또한 명령형 어미의 경우 대부분 문장의 마지막 어절로 나타난다. 따라서 분석결과 명령형 어미로 분석이 되는 경우, 다른 형태의 분석 후보의 우선순위를 높이고, 또한 문장의 마지막 어절로 사용되지 않는다면 분석 후보에서 삭제가 가능하다.

2. 후행 어절에 따른 결합 제약

가. 어미 '-지'의 후행 어절

어미 '-지'는 특정한 후행 어절을 요구하는데 그것은 '말다, 않다, 못하다' 등이다. 따라서 '-지'가 어미로 쓰인 경우 후행하는 어절을 조사하여 이를 충족하면 어미로 쓰일 수 있으며 그렇지 않으며 어미로 쓰일 수 없는 어절이다.

다음의 예는 '-지'가 어미로 분석 될 경우, 후행 어절에 따라 중의성을 제거 할 수 있음을 보여준다.

(3-6) 여러 가지 질병에 대한 설명을 들었다.
(3-7) 나는 오늘 학교에 가지 않았다.

(3-6R) 가지(체언) 질병(체언)+에(조사)
*가(용언)+지(어미)
(3-7R)*가지(체언) 않(용언)+았(선어미)+다(어미).
가(용언)+지(어미)

3. 선행 어절에 따른 결합 제약

가. '-ㄹ 수가 있/없다'의 선행 어절

'-ㄹ 수가 있/없다'와 같은 복합 어절은 선행 어절로 용언을 요구한다. 이러한 선행 어절에 대한 제한을 통해 중의성 해소를 할 수 있으며 다

음은 그 예를 보인다.

(3-8) 궁금증은 참을 수가 없었다.
(3-9) 모기가 물어 대면 밤새 잘 수가 없다.
(3-8R) 참(용언)+음(어미)
*참(체언)+을(조사)
수(체언)+가(조사) 없(용언)+었(선어미)+다(어미)
(3-9R) 자(용언)+ㄹ(어미)
*잘(독립언)
*잘(체언)
수(체언)+가(조사) 없(용언)+었(선어미)+다(어미)

4. 후처리를 통한 가중치 부여

조사/어미의 결합 제약, 선행 어절에 의한 결합 제약, 또는 후행 어절에 의한 결합 제약에 의해서도 중의성이 해결되지 않는 어절에 대해서는 분석 후보의 유형에 따라 절대적 가중치 부여방법과 상대적 가중치 부여 방법을 이용하여 각 분석 후보들 간의 우선 순위를 부여한다. 이러한 방법은 각각의 분석 후보 유형에 따른 통계적 정보에 의해 이루어질 수 있다.

Ⅳ. 실험 및 결과

본 논문에서 제안된 형태소 분석시 발생하는 중의성 해소를 위한 형태소 분석기를 Workstation(AlphaStation 255)에서 C언어로 구현하였으며, 기존의 형태소 분석기에 중의성 축소 기능을 부여한 형태의 시스템이다. 사전은 뉴에이스 국어사전[9]의 내용을 형태소 분석을 위한 사전으로 구축하여 사용하였으며 전체 12만 어휘가 등록되어 있다.

실험에 사용된 자료는 한국과학기술원의 대한민국 국어정보베이스[10]의 말뭉치와 품사 태깅된 자료의 일부를 이용하였다.

중의성 해소의 판단은 원시 말뭉치를 구현된 형태소 분석기로 분석한 결과, 하나의 분석 결과가 나오고 그 결과가 품사 태깅된 결과와 일치되는 경우와, 분석된 후보가 여러개일 경우 가장 우선 순위가 높은 후보가 품사 태깅된 결과와 일치하는 경우 중의성이 해소되었다고 본다. 중의성 해소의 비율은 이전의 형태소 분석기[11]에서 중의성이 발생하는 단어에 대해서 새로 제안된 형태소 분석기의 결과에서 중의성 해소의 비율로 구한다.

형태소 분석시 중의성이 발생하는 어절은 말뭉치에 따라 다르지만 평균 20%에 해당한다. 표 1은 전체어절에 따른 중의성 발생 어절을 나타낸다. 이 중의성 발생 비율은 사용하는 형태소 분석기에 의해서도 달라질 수 있으며, 본 실험에서는 범용 한국어 형태소 분석기[11]를 사용하였다.

표 2는 표 1에서 중의성이 발생한 어절에 대해 본 논문에서 제안된 형태소 분석기를 통하여 분석한 결과 중의성이 제거된 비율을 나타낸다.

표 1. 중의성 발생 어절

구 분	전체대상 어절	중의성이 있는 어절	중의성이 발생한 어절	중의성이 발생하지 않는 어절
말뭉치 1	7284	4605 (63.2%)	1264 (17.4%)	1415 (19.4%)
말뭉치 2	9900	6305 (63.7%)	2285 (23.1%)	1310 (13.2%)
계	17184	10910 (63.5%)	3549 (20.6%)	2725 (15.9%)

표 2에서 보듯이 형태소 분석 과정에서 생기는 형태론적 중의성은 75% 이상이 완전하게 제거되며, 분석 후보가 두 개 이상인 경우에는 각각의 후보 유형별로 구분하여 가중치를 부여하여 우선 순위가 반영되어 출력된 것까지 고려하면 92%이상의 중의성이 해소됨을 볼 수 있다. 그러나 중의성이 해소되지 않은 어절들도 8%이상이나 되었다.

표 2. 중의성 해소의 비율

구 분	중의성 발생어절	중의성 해소어절	기본원칙에 따른 중의성 해소어절	중의성이 해소되지 않은어절
말뭉치1	1264	989 (78.2%)	187 (14.8%)	88 (7%)
말뭉치2	2285	1672 (73.2%)	417 (18.2%)	196 (8.6%)
계	3549	2661 (75%)	604 (17%)	284 (8%)

표 3은 중의성 해소에 실패한 어절들에 대한 분석 결과이며, 각 유형별로 중의성이 해소되지 않은 어절들에 대한 예를 보인다.

표 3. 중의성 해소가 안된 어절 분석

중의성 유형	중의성 발생 원인	중의성 발생 어절
어근 유형 중의성	X(체언)+은/는/을(조사) X(용언)+은/는/을(어미)	일을, 계시는, 거하는, 길을, 죄는, 적은, 구하는, 되는, 우리는, 남을, 하는, 이는, 여기는, 배우는, 들은, 데는. 등
	X(체언)+하고(조사) X(체언)+하(접사)+고(어미)	간구하고, 감사하고, 기도하고, 기초하고, 증가하고, 탄원하고, 파괴하고, 등
	X(체언)+라고(조사) X(용언)+라고(어미)	죄라고, 구하라고, 등
원형 복원 중의성	길을: 길(용언)+을(어미) 긴(용언)+을(어미)	길을, 들은, 갈, 간, 건, 팔, 살, 등
형태소 분리 중의성	보기로: 보기(체언)+로(조사) 보(용언)+기로(어미)	보기로, 그리고, 임의, 자본주의, 등

중의성 해소가 안 된 어절들은 대부분이 어근 유형 중의성에 해당되는 어절들이었다. 특히 문법 형태소 '-은/-는/-을'은 조사와 어미 모두로 쓰이며 매우 빈번히 사용된다. 이러한 문법 형태소와 결합하는 어휘 형태소가 체언-용언의 품사 중의성을 갖는 단어들은 중의성을 발생하며 그로 인해 발생하는 중의성은 대부분 해소되지 않았다.

표 3에서 보인 어절들에 대한 중의성 해소는 어절 자체만으로의 형태론적인 특성만을 이용하여 해결하는 것은 불가능하며, 선행어절 또는 후행어절로부터의 결합 제약조건도 없을 경우에는 형태소 분석시 발생하는 중의성은 해결할 수 없다. 이러한 중의성은 상위 분석 단계인 구문 분석이나 의미 분석에서 해소될 수 있다.

V. 결 론

본 논문에서는 한국어 형태소 분석시 발생하는 중의성 중 어미/조사 결합시 발생하는 중의성 해소를 위하여 어절 자체에서 또는 선행 어절이나 후행 어절에 따른 결합 제약을 이용하였다. 또한 형태소 분석기에서 특별한 처리를 필요로 하는 따옴표와 묶음표에 대한 올바른 처리 방안과 한영 혼용어에 대한 처리 방안을 형태소 분석의 전처리 기능으로서 제시하였다.

중의성이 발생하는 어절 중 평균 75%에 해당하는 어절이 조사/어미의 결합 제약 강화로 해소되었다. 또한 후처리를 통한 가중치를 이용하여 전체 92%의 중의성을 해소할 수 있었다.

한국어 형태소 분석시 발생하는 중의성의 유형 및 수는 매우 다양하며, 각 유형의 발생 원인과 체계적인 분류와 함께 각각의 중의성 해소 방법에 대한 연구가 향후 연구 과제이다.

참 고 문 헌

[1] 임권묵, 김병희, 송만석, "형태 중의성 해결을

위한 말마디 사전 설계", 한국정보과학회 봄 학술발표 논문집, 제 20권 1호, pp.789-792, 1993.

[2] 안미정, 옥철영, "형태소적 중의성 해소를 통한 한국어 복수동사의 통합", 한국 인지과학회 춘계 학술발표 논문집, pp.18-31, 1995.

[3] 임희석, 이호, 임해창, "형태소 분석 단계에서 발생하는 어절의 중의성 분석방안", 한국정보과학회 봄 학술발표 논문집, 제 20권 1호, pp.773-776, 1993.

[4] 김충원, 임권묵, 송만석, "의미정보를 이용한 형태소 중의성 해결", 한국정보과학회 가을 학술발표 논문집, 제 21권 2호, pp.649-652, 1994.

[5] 이하규, "어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소", 정보과학회 논문지 (B), 제 24권 1호, pp.82-89, 1997.

[6] 김영훈, 정천영, 김남철, 왕지현, 서영훈, "Viterbi 알고리즘을 이용한 효율적인 품사 태깅", 한국정보과학회 봄 학술발표 논문집, 제 23권 1호, pp.969-972, 1996.

[7] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사 태깅", 정보과학회 논문지 제 22권 1호, pp.136-146, 1995.

[8] 이희승, 안병희, 고친판 한글 맞춤법 강의, 신구문화사, 1995.

[9] 운평어문연구소, 뉴에이스 국어사전, 금성교과서주식회사, 1994.

[10] 최기선, "대한민국 국어정보베이스 - Evaluation Release 0.1 (Korea National Language Information Base)", 한국과학기술원, 1997.

[11] 장동수, "음절에 기반한 한국어 형태소 분석기", 충북대학교 대학원 컴퓨터공학과 석사학위 논문, 1994.