접속 특성과 말마디 사전을 이용한 형태소 분석

임 권 묵† 송 만 석††

요 약

본 논문은 형태소의 접속 특성과 대형 말뭉치(corpus)로부터 추출된 중의성 말마디의 인접 정보를 이용해서 한국어 형태소 분석기를 구현한다. 일반적으로 말마디는 형태소의 접속 특성과 결합 규칙을 적용함으로써 하나의 결과로 분석될 수 있으나 중의성 말마디는 가능한 결과들로부터 적절한 하나를 선택하기 위해서 인접 말마디 정보나 문법 정보 또는 문맥 정보 등이 요구된다. 그러나 문법 정보와 문맥 정보는 구문 분석과 의미 분석 단계를 거쳐야만 가능하기 때문에 여기서는 표층적인 정보로서 인접 말마디 정보를 이용한 중의성 해결을 시도하였다. 형태소의 접속 특성과 중의성 말마디의 인접 정보를 사전에 수록함으로써 축약어와 불필요한 결과를 제시하는 말마디 그리고 중의성 말마디까지도 형태소 분석이 거의 가능하게 된다. 본 분석기의 효능은 정확하고 풍부한 정보를 사전에 효율적으로 수록함으로써 이룩될 것이며, 이를 위해 형태소 사전과 말마디 사전을 데이타베이스로 설계하고, 필요한 정보들을 대형 말뭉치로부터 추출하여 사전에 저장한다.

Morphological Analysis with Adjacency Attributes and Phrase Dictionary

Kwon-Mook LIM† and Man-Suk SONG††

ABSTRACT

This paper presents a morphological analysis method for the Korean language. The characteristics and adjacency information of the words can be obtained from sentences in a large corpus. Generally a word can be analyzed to a result by applying the adjacency attributes and rules. However, we have to choose one from the several results for the ambiguous words . The collected morpheme's adjacency attributes and relations with neighbor words are recorded in a well designed dictionaries. With this information, abbreviated words as well as ambiguous words can be almost analyzed successfully. Efficiency of morphological analyzer depends on the information in the dictionaries. A morpheme dictionary and a phrase dictionary have been designed with lexical database, and necessary information extracted from the corpus is stored in the dictionaries.

1. 서 론

형태소란 자연언어에서 '의미를 갖는 최소 단위'를 말하며, 자연어 처리에서의 형태소 분석은 주로 형태론적 변형 현상을 처리하는 것이다[1]. 형태론적 변형은 이웃하는 형태소끼리의 결합 조건에 의해 형태소의 일부가 교체되거나 삽입 혹은 삭제되는 현상으로 한국어 용언의 불규칙 현상이 이에 속한다[2]. 그러나 한국어나 중국어와 같이 여러 개의 형태소가 하나의 단어를 이루는 언어에서는 형태론적 변형보다 말마디를 구성하고 있는 형태소들을 분리하는 일이 더 중요한 문제로 인식되고 있다[3,4].

형태소 분석은 자연어 처리에서 구문 분석이나 의미 분석의 전(前)단계일 뿐만아니라 기계번역이나 정보검색 등 모든 자연언어 관련분야에서 수행되어야 할 필수적 과정으로서 매우 중요하다. 특히 한국어와 같이 단어의 통사적-의미적 기능이 어순보다 조사나 어미와 같은 형식 형태소에 의해 결정되는 언어에서는 형태소 분석이 구문 분석과 의미

†정 회 원 : 대신대학교 전자계산학과 조교수
††정 회 원 : 연세대학교 전산과학과 교수
논문접수 : 1994년 2월 16일, 심사완료 : 1994년 4월 25일

분석에 미치는 영향이 매우 크기 때문이다[5]. 그러나 중요성에 비해서 한국어 형태소 분석의 연구가 지금까지 부진했던 이유는 영어와 같은 인구어의 형태소 분석이 한국어에 비해 매우 간단해서 형태소 분석을 당연한 것으로 간주하고 그동안 구문 분석이나 의미 분석에만 주로 관심을 두었기 때문이다[6, 7].

현재까지 연구된 한국어 형태소 분석법에는 여러 가지가 있으나 그 중에서 대표적인 것으로는 Two-Level 형태론, Tabular Parsing 방법, 최장일치법 등이 있다[4, 8, 9]. 모든 언어에 공통적으로 적용되는 Two-Level 형태론은 입력 문자열로부터 각 형태소들의 원형을 추정하는 규칙을 유한단계자동으로 표현 처리하는 방법이다. 이방법은 굴절이 심하게 일어나는 언어의 형태소를 분리하고 원형을 복구하는데는 효율적이나[8, 9] 한국어와 같이 언어의 문법적 기능이 어근과 접사의 결합 연속에 의하여 나타나는 교착어에서는 그 규칙의 수가 계속 증가하게 되므로 많은 규칙을 기술해야 하는 문제가 있다. Tabular Parsing 방법은 길이가 n인 말마디에 대해 i 번째 자소로부터 j 개의 자소로 구성되는 형태소를 (i,j)쌍으로 표시하여 크기가 (n×n)/2인 삼각테이블을 구성하고 (1,n)쌍을 구성하는 형태소들을 추출하는 방법이다. 그러나 이 방법은 접속 정보표를 구성하기가 어렵고 사전탐색 횟수가 많아 실행속도가 느려진다[4]. 최장일치법은 말마디를 가능한 모든 형태소로 분할한 다음 그 말마디를 이루고 있는 형태소들의 집합 중에서 가장 긴 형태소를 포함하는 것을 우선적으로 검사하여 선택하는 단순하고 명쾌한 방법이나 중의성을 해결하지 못하는 단점이 있다.

최근에는 자소 단위가 아닌 음절 단위의 형태소 분석이 주로 연구되고 있으며, 앞에 열거한 방법들의 단점을 보완하거나 시간을 절약하는 알고리즘들도 제시되고 있다[10]. 대표적인 것으로는 조사나 어미의 첫음절부터 끝음절까지 각 위치별로 음절사전을 구성하여 분리를 시도하는 방법과 부분적으로 구문 분석이 이루어진 가운데 말마디 중의성을 해결하는 방법이 제시되었다[11].

본 논문은 형태소 분석을 자연어 처리의 가장 기초적이고 필수적인 단계로 인식하고, 구문 분석이나 의미 분석이 수행되기 전(前)단계로서 가능한 많은 말마디들의 정확한 형태소 결과를 제공함으로써 다음 단계에서 수행해야 할 작업량을 줄인다. 또한 형태소 분석 단계에서 처리하기 어려운 축약어나 분석에 실패한 말마디 그리고 복수 결과를 보여 주는 말마디들의 분석결과를 사전에 수록하고 탐색함으로써 처리 시간을 현저히 줄인다. 따라서 본 논문은 대형 말뭉치를 이용하여 한국어의 특성을 분석하여 한국어에 적합한 형태소 분석기를 구현하고, 분석에 실패한 말마디와 중의성 말마디를 인접 정보와 함께 말마디 사전에 수록한다.

2. 한국어의 특징

인위적으로 만들어진 형식언어와는 달리 자연언어는 모호성이 많이 발견되고 예외규칙이 많으며 언어가 시대의 흐름에 따라 변한다는 공통적인 특징을 갖고 있다. 이 외에도 각각의 언어는 서로 다른 특성들을 갖고 있는데 한국어의 특성으로는 알타이 어족으로서 모음조화 현상을 들 수 있고 한국어에만 있는 특성으로는 음절 단위로 표기하는 것과 띄어쓰기 규칙이 있다[12].

2.1 음절과 자소

일반적으로 자연언어를 분석하는 단위는 자소로서 영어와 같은 인구어는 여러 개의 알파벳이 모여 하나의 말마디를 이루지만 한국어의 글자 체계는 자음(ㄱ, ㄴ, ㄷ, …)과 모음(ㅏ, ㅑ, ㅓ, …)이 결합되어 열린음절과 닫힌음절을 이루고 다시 여러 개의 음절이 모여 하나의 말마디를 이룬다. 영어와는 달리 한국어에서는 말마디를 구성하는 자음과 모음이 심한 제약을 갖는다. 초성은 단자음 하나 또는 음가가 없는 자음 'ㅇ'으로 이루어지고, 중성은 모음하나로 이루어지며, 종성은 단자음과 복자음으로 이루어진다.

말마디 ::= { 음절 }*
음절 ::= 열린음절 | 닫힌음절

열린음절 ::= 초성 + 중성
닫힌음절 ::= 초성 + 중성 + 종성

초성, 중성, 종성을 구성하는 문자의 수는 초성이 19개의 자음, 중성이 21개의 모음, 종성이 27개의 자음으로 구성되므로, 가능한 모든 음절의 수는 11,172 (19 * 21 * 28)개이다.

2.2 말마디와 형태소

한국어의 말마디는 음절 단위로 표기하고 각 말마디는 띄어 쓰는 것이 원칙이므로 말마디의 구분은 매우 용이하다. 따라서 한국어 형태소 분석은 실질적으로 입력된 말마디를 분석하는 것으로부터 시작된다. 영어와 같은 굴절어는 어근에 접두사나 접미사가 결합되고 다시 굴절이 일어나는 규칙에 따라 말마디가 구성되나 한국어와 같은 교착어는 굴절어가 가지는 속성 외에도 여러 개의 형태소가 결합하여 하나의 말마디를 이룬다. 말마디를 이루는 형태소들은 공통된 성질에 따라 실질 형태소와 형식 형태소로 분류되며, 이것은 <표 1>과 같이 세분화된다. 본 논문에서는 학교 문법체계인 9 품사를 기본으로 전산처리에 적합하도록 어미와 접미사를 세분하여 첨가하였다.

<표 1> 형태소 분석을 위한 품사 분류

<Table 1> Part of Speech for Morphological Analysis

실질형태소	체 언	명사, 대명사, 수사
	용 언	동사, 형용사, 존재사, 보조동사, 서술격 조사
	수식언	관형사, 부사
	독립언	감탄사
형식형태소	관계언	조사
	어 미	어말 어미, 선어말 어미, 명사형 전성어미
	접미사	체언 접미사, 부사화 접미사, 용언화 접미사

2.3 말마디의 구조

한국어의 실질적인 형태소 분석은 띄어쓰기로 구분되어 있는 말마디들을 형태소로 분리하여 원형으로 복원하는 과정이므로 말마디의 구조에 관해서 자세히 분석해 본다. 한국어 말마디는 하나 이상의 형태소로 구성되어 있으며, 구성 전이도로서 (그림 1)과 같이 나타낼 수 있다[13, 14].

(1) 서술격 조사는 체언의 우측에 결합될 수 있다는 점에서 조사의 성격을 갖고 있다. 또한 체언류에 붙어 말마디를 용언의 성격으로 바꾸며 우측으로 어미와 접속하는 점에서는 용언화 접미사와 같은 성질을 갖고 있다.

(2) 조사는 어근 형태소의 종성에 따라 결합이 달라지며, 보조사는 부사와도 결합이 가능하다. 조사와 조사가 겹쳐서 하나의 조사를 이루는 복합조사는 하나의 조사로 처리한다.

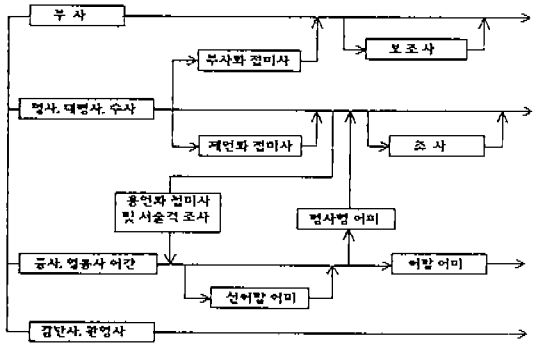
(3) 어말 어미는 어근 형태소가 갖고 있는 종성의 종류, 품사에 따라 결합하는 어미가 달라진다. 특수한 경우로서 어미 뒤에 일부 조사가 결합될 수 있으나 이를 따로 처리하지 않고 결합형태를 하나의 어미로 처리하였다.

(4) 선어말 어미는 시제, 높임 선어말 어미만을 처리한다. 선어말 어미끼리의 결합은 복합 선어말 어미로 사전에 수록하여 처리하며, '시겠', '시었', '시었겠' 등이 있다.

(5) 명사형 전성어미는 어말 어미와는 달리 용언 어간의 우측에 결합하여 말마디를 체언의 성격으로 바꿔 주는 역할을 하며, 우측에 조사와 결합할 수 있는 특징을 갖고 있다.

(6) 체언 접미사는 체언 어근의 우측에 접속하여 말마디의 성격을 체언으로 유지하는 성질을 갖고 있다.

(7) 부사화 접미사는 체언 어근의 우측에 접속하여 말마디를 부사의 성격으로 바꾸는 성질을 갖고 있으며, '스레', '없이' 등을 말한다.



(그림 1) 말마디의 구성 전이도

(Fig. 1) Transition Diagrams of a Word

(8) 용언화 접미사는 체언이나 용언 어근의 우측에 접속하여 말마디를 용언의 성격으로 바꾸는 성질을 갖고 있다.

2.4 말마디의 분류

본 논문에서는 말마디를 전산처리하는 기준에 따라 접속 특성으로 분석이 용이한 말마디, 중의성은 없으나 접속 특성으로 분석이 어려운 말마디, 중의성을 갖는 말마디로 분류한다[15].

(1) 접속 특성으로 분석이 용이한 말마디

어휘적 접속 특성과 통사적 접속 특성만으로 형태소 분석이 가능한 말마디이다.

(2) 중의성은 없으나 접속 특성으로 분석이 어려운 말마디

① 의미적 접속 특성의 부재로 인한 말마디-의미적인 접속 특성이 사전에 수록되지 않아 분석에 실패하는 말마디이다.

(예) 얼마 ——> 얼(동사)+마(어미)

위 예는 의미적 접속 특성의 부재로 인하여 발생하는 분석 오류이다.

② 약어-조사나 어미가 축약되어 접속 특성으로 분석이 어려운 말마디이다.

(예) 난 ——> 나(대명사)+는(조사)

(3) 중의성을 갖는 말마디

한국어 말마디가 가질 수 있는 중의성에는 크게 3 가지 경우로서, 형태 중의성, 품사 중의성 그리고 의미 중의성이 있다.

① 형태 중의성-하나의 말마디가 서로 다른 2 가지 이상의 형태소 유형으로 해석되는 경우에 발생하는 중의성이다.

(예) 종이 ——> 종이(명사)

종 (명사)+이 (조사)

② 품사 중의성-말마디로부터 분리된 하나의 형태소가 서로 다른 여러 개의 품사로 해석되는 경우에 발생하는 중의성이다.

(예) 안은 ——> 안(명사)+은(조사) : inside

안(동사)+은(어미) : embrace

③ 의미 중의성-같은 형태소와 품사를 갖는 동일한 형태소가 문장 내에서 서로 다른 의미로 해석되는 경우에 발생하는 중의성이다.

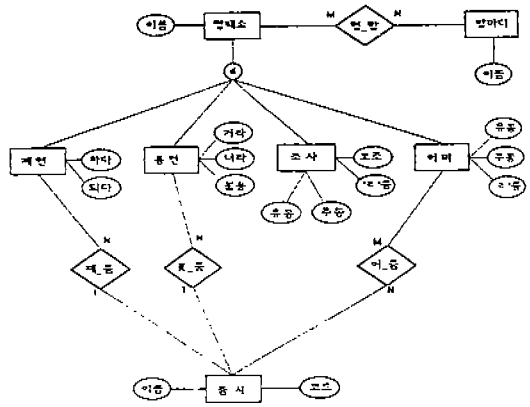
(예) 달려 ——> 달리(동사)+어(어미) : hang

달리(동사)+어(어미) : run

3. 사전의 구성

기존의 자연어 처리는 대부분 문맥자유문법(Context-Free Grammars) 계열의 문법 규칙들을 그대로 적용하였으며, 이것은 모든 언어 현상이 어떤 일정한 규칙이나 법칙에 의하여 설명될 수 있다는 가정에서 시도된 것이다[16]. 그러나 실제로 자연언어는 형식언어와는 달리 규칙에 의해서만 해결될 수 없는 예외적인 사항을 많이 내포하고 있을 뿐만아니라 각각의 언어는 나름대로의 서로 다른 특성들을 안고 있다. 따라서 최근의 형태소 분석은 규칙에만 의존하지 않고 사전을 기반으로 하는 방향으로 많이 연구되고 있다[10, 11, 14, 15]. 본 논문은 말뭉치를 통해 분석된 어휘정보를 기반으로 사전을 구성하여 형태소분석을 시도한다.

3.1 Entity-Relationship 모델



(그림 2) Entity-Relationship 모델

(Fig. 2) Entity-Relationship Model

Entity로는 의미를 제외한 말마디, 형태소, 품사

등이 있으며, 형태소는 다시 subclass로 체언, 용언, 조사, 어미 형태소로 나누어진다. Entity들간의 relationship은 각 형태소들과 품사 사이에 cardinality ratio N:1을 갖는 relationship들이 있고, 형태소와 말마디 사이에 M:N의 cardinality ratio를 갖는 relationship이 존재한다(그림 2).

3.2 형태소 사전

사전은 기능어 사전(function word)과 체언류 사전, 용언류 사전으로 구성된다. 기능어 사전은 조사, 어미, 선어말 어미, 체언 접미사, 용언화 접미사, 부사화 접미사 사전으로 나누고, 각 사전은 기능어와 각 기능어들의 결합정보를 갖고 있다. 기능어 사전은 분석기에서 자주 사용되고 용량이 작으므로 주기억 장치에 저장 사용한다. 체언 사전과 용언 사전은 보조 기억 장치에 저장하고 인덱스를 이용하여 해당 부분을 탐색한다[17]. 각 사전의 형태소 정보는 한국어 사전의 표제어를 추출하여 수록하였다[18].

(1) 체언 사전-체언 사전에는 체언류 205,701개(명사 193,341개, 대명사 211개, 부사 10,263개, 감탄사 629개, 수사 161개, 관형사 1,096개)를 품사정보와 함께 수록하였다. 명사의 경우에는 용언화 접미사 '하다', '되다'와의 결합이 가능한 가를 나타내는 정보를 함께 수록하였으며, 명사와 명사가 결합해서 이루어진 복합명사는 하나의 명사로서 수록되었다.

(체언형태소, 품사, '하'결합, '되'결합)

(2) 용언 사전-용언 사전에는 용언류 20,642개(동사 15,027개, 형용사 5,615개)를 품사정보, 불규칙 정보와 함께 수록하였다. 불규칙 정보는 '거라', '너라'와의 결합 정보와 불규칙 용언 정보가 있다.

(용언형태소, 품사, 거라, 너라, 불용언)

(3) 조사 사전-말뭉치로부터 실제 사용되는 조사 350개를 추출하여 각각의 결합정보와 함께 조사 사전에 수록하였다. 결합 정보는 조사 형태소와 부사와의 결합 관계와 유종성('ㄹ'제외), 무종성, 'ㄹ'종성 형태소와의 결합 관계를 나타낸다.

(조사형태소, 보조사, 유종성, 무종성, 'ㄹ'종성)

(4) 어미 사전-말뭉치로부터 실제 사용되는 어미 820개를 추출하여 각각의 결합정보와 함께 어미 사전에 수록하였다. 결합 정보는 각 어미 형태소와 선어말 어미(선말), 동사, 형용사, 존재사(있다, 없다 등), 지정사(-이다) 등과의 결합 관계와 함께 유종성('ㄹ' 제외), 무종성, 'ㄹ'종성 형태소와의 결합 관계를 나타낸다.

(어미형태소, 선말, 동사, 형용, 존재, 지정, 유종, 무종, 'ㄹ'종성)

(5) 기타 사전-기능어 사전으로 체언 접미사 11개('전', '끼리', '들' 등), 용언화 접미사 9개('답다', '당하다', '되다' 등), 보조용언 23개('가다', '만하다', '못하다' 등)를 각각 수록했다.

3.3 말마디 사전

중의성을 갖는 말마디 582개('가지', '갈', '날', '단' 등)와 접속 특성으로 분석이 어려운 말마디 525개('순간', '시간', '알아', '언제나' 등)를 수록했다. 접속 특성을 이용하는 경우에 처리하는 데 많은 시간을 소모하는 말마디들을 말마디 사전에 입력함으로써 형태소 분석의 효율성을 이룰 수 있다. 본 논문에서는 길이가 5 음절 이상으로 자주 사용되는 말마디들을 말마디 사전에 입력하여 처리 시간을 단축하였다. 후에 사전 표제어의 증가로 인하여 전체적인 분석의 효율이 감소하는 경우에는 효율에 따라 등록 표제어의 수를 줄인다.

말마디 사전은 특성 화일, 유형 화일, 정보 화일로 크게 나누어진다. 특성 화일은 각 말마디의 특성과 중의 특성을 코드로서 표현하고, 유형 화일은

각 말마디가 갖을 수 있는 모든 형태소 분석 유형을 제시하며, 정보 화일은 중의성을 해결하기 위한 인접 정보와의 관계를 나타낸다[19].

(1) 말마디특성 화일-말마디와 코드의 attribute로 구성되며, 중의성을 갖는 모든 말마디와 각 말마디의 특성을 코드로서 수록한다. 말마디의 코드는 형태 중의성과 품사 중요성의 유무와 함께 각각의 중의성을 해결하기 위해 찾는 정보의 수를 표기하였다.

말마디특성(말마디, 코드)

(2) 형태중의특성 화일-말마디, 색인, 특성의 attribute로 구성되며, 형태 중의성을 갖는 모든 말마디와 각 말마디의 중의성을 해결하기 위한 정보의 위치와 종류를 표시한 특성을 함께 수록한다. 각 말마디는 특성의 수에 따라 색인을 갖으며, 말마디와 색인의 결합키(Combination Key)로서 특성을 찾아간다.

형태중의특성(말마디, 색인, 특성)

(3) 형태유형 화일-말마디, 색인, 형태소, 품사의 attribute로 구성되며, 형태 중의성을 갖는 모든 말마디와 각 말마디가 분석되는 유형별로 형태소, 품사, 빈도율을 포함한다. 이 테이블은 말마디의 중의성이 해결되었을 때 해당되는 결과를 찾거나 분석에 실패한 말마디의 가능한 유형 전부를 제시할 때 사용된다.
형태유형(말마디, 색인, 형태1, 품사1, 형태2, 품사2, 형태3, 품사3, 빈도)

(4) 형태중의정보 화일-말마디의 형태 중의성을 해결하기 위한 인접 정보를 갖고 있다.

형태중의정보(말마디, 인접정보, 색인)

(5) 품사중의특성 화일-형태소, 색인, 특성의 attribute로 구성되며, 품사 중의성을 갖는 모든 형태소와 각 형태소의 중의성을 해결하기 위한 정보의 위치와 종류를 표시한 특성을 함께 수록한다. 각 형태소는 특성의 수에 따라 색인을 갖으며, 형태소와 색인의 결합키(Combination Key)로서 특성을 찾아간다.

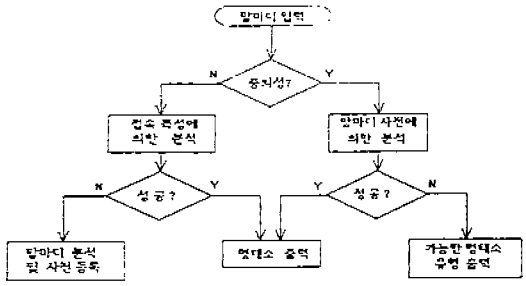
품사중의특성(형태소, 색인, 특성)

(6) 품사중의정보 화일-형태소의 품사 중의성을 해결하기 위한 인접 정보를 갖고 있다.

품사중의정보(형태소, 인접정보, 품사)

4. 형태소 분석기의 설계

접속 특성만으로 분석이 가능한 말마디와 중의성을 갖는 말마디를 분류하여 형태소 분석을 실행한다(그림 3).



(그림 3) 형태소 분석 절차
(Fig. 3) Procedure of Morphological Analysis

(1) 말마디 입력-문장 단위로 말마디들을 입력하여, 각 말마디를 말마디 사전으로부터 탐색하여 등록되어 있는가를 확인한다.
(2) 접속 특성에 의한 분석-접속 특성을 이용한 분석 방법으로서 말마디 사전에 등록되지 않은 말마디의 형태소 분석을 실시한다.
(3) 말마디 사전에 의한 분석-말마디 사전을 이용한 분석 방법으로서 말마디 사전에 등록되어 있는 말마디의 형태소 분석을 실시한다.
(4) 말마디분석 및 사전등록-말마디 사전에 등록되지 않음으로써 분석에 실패한 말마디를 분석하여 말마디 사전에 등록한다.
(5) 형태소 출력-말마디 사전이나 접속 특성을 이용하여 분석에 성공한 결과를 출력한다.
(6) 가능한 형태소 유형 출력-말마디 사전에 등록

되어 있으나 분석에 실패한 말마디의 가능한 모든 유형을 제시한다.

4.1 접속 특성에 의한 분석

접속 특성을 이용한 형태소 분석은 사전에 형태소와 그 형태소의 좌우 접속 특성을 저장하고, 이 접속 특성을 이용하여 형태소들의 접속 가능 여부를 검사한다. 형태소 분석은 말마디의 우측에서부터 좌측 방향으로 종성자소 혹은 음절 하나씩을 분리하면서 기능어들을 먼저 처리한다[17]. (그림 4)에 분석 절차를 전이도로 보인다.

(1) 체언 탐색 및 확인-입력어(모듈에 입력된 문자열)를 체언사전으로부터 탐색하여 체언 형태소라는 것을 확인한다. 확인된 체언 형태소는 분석 결과로서 출력되고, 입력어는 체언접미사 분리를 시도한다.

(2) 체언 접미사 분리-입력어의 우측부터 음소를 하나씩 분리하여 체언 접미사임을 확인한다. 체언 접미사가 제거된 어간은 체언 탐색 및 확인을 시도하고, 입력어는 명사형 어미 분리를 시도한다.

(3) 명사형 어미 분리-입력어의 우측부터 음소를 하나씩 분리하여 명사형 어미임을 확인한다. 명사형 어미가 제거된 어간은 선어말 어미 분리를 시도하고, 입력어는 조사 분리를 시도한다.

(4) 조사 분리-입력어의 우측부터 음소를 하나씩 분리하여 조사임을 확인한다. 조사가 제거된 어간은 체언 탐색 및 확인을 시도하고, 입력어는 어미 분리를 시도한다.

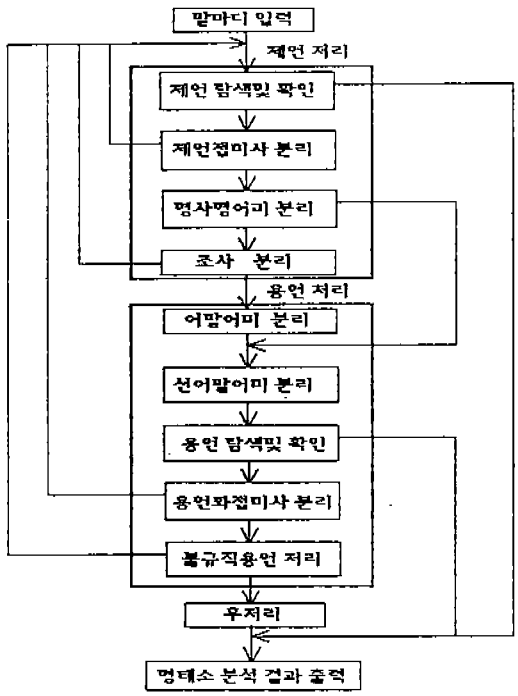
(5) 어말 어미 분리-입력어의 우측부터 종성자소 혹은 음소를 하나씩 분리하여 어미임을 확인한다. 어미가 제거된 어간은 선어말 어미 분리를 시도한다.

(6) 선어말 어미 분리-입력어의 우측부터 종성자소 혹은 음소를 하나씩 분리하여 선어말 어미임을 확인한다. 선어말 어미가 제거된 어간은 용언 탐색및 확인을 시도한다.

(7) 용언 탐색 및 확인-입력어를 용언사전으로부터 탐색하여 용언 형태소라는 것을 확인한다. 확인된 용언 형태소는 결과로서 출력되고, 입력어는 용언화 접미사 분리를 시도한다.

(8) 용언화 접미사 분리-입력어의 우측부터 음소를 하나씩 분리하여 용언화 접미사임을 확인한다. 용언화 접미사가 제거된 어간은 체언 탐색및 확인을 시도하고 입력어는 불규칙 용언 처리를 시도한다.

(9) 불규칙 용언 처리-어간 변화형 불규칙 용언('ㄷ', 'ㅅ', 'ㄹ', 'ㅎ')과 어미 변화형 불규칙 용언('여', '거라', '너라', '러', '르', 'ㅂ', '으', '우')을 처리한다.

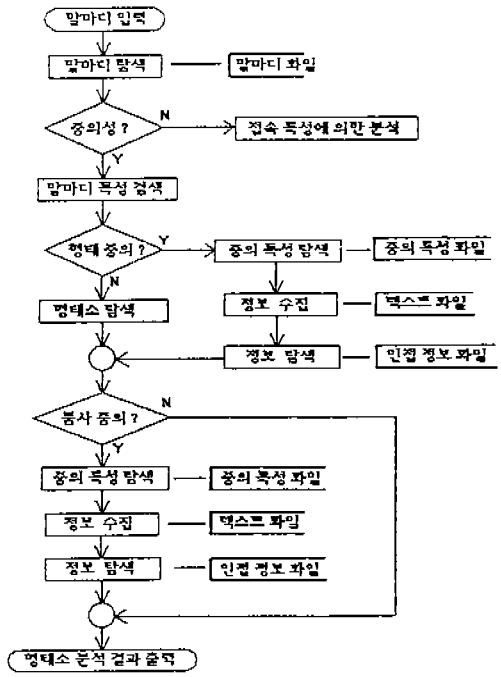


(그림 4) 접속 특성을 이용한 분석 절차
(Fig. 4) Analysis Procedure with Adjacency Attributes

(10) 후처리-용언의 어간과 어미 사이에 일어나는 음운 축약 현상을 원형 복원하여 처리한다. ('돼'->'되어', '치'->'하지' 등)

4.2 말마디 사전을 이용한 분석

말마디사전을 이용한 형태소 분석은 말마디 정보와 형태소 분석결과를 사전에 미리 수록하고, 인접 말마디로부터 수집된 정보를 이용하여 형태소 분석 결과를 추출해 낸다. 이 방법은 우리가 이미 알고 있는 말마디의 분석결과를 데이타베이스를 이용해 설계한 사전에 효율적으로 저장함으로써 쉽게 결과를 얻을 수 있다. 또한 규칙으로는 해결할 수 없는 말마디나 중의성을 갖는 말마디는 좌우 인접 말마디의 정보를 참조함으로써 쉽게 분석결과를 얻을 수 있다(그림 5). 여기서 사용되는 인접 말마디의 정보는 대형 말뭉치를 이용한 용례 분석을 통하여 수집한다[15, 19].



(그림 5) 말마디 사전을 이용한 분석 절차
(Fig. 5) Analysis Procedure with Phrase Dictionary

(1) 말마디 특성 검색-말마디 특성 테이블로부터 입력된 말마디의 특성을 조사해 중의성 여부를 판단한다.
(2) 형태소 탐색-중의성이 없는 말마디의 형태소 갯수에 따라 테이블에서 형태소 분석 결과를 읽어온다.
(3) 중의 특성 탐색-중의 특성 테이블로부터 중의성을 갖는 말마디의 특성을 조사해 참조할 정보의 위치를 판별한다.
(4) 참조할 정보 인식-말마디의 중의 특성을 분석하여 문장으로부터 참조할 정보가 무엇인가를 인식한다.
(5) 문장으로부터 정보 탐색-참조할 정보를 문장으로부터 탐색한다.
(6) 정보에 의한 형태소 탐색-말마디와 참조한 정보로서 해당 테이블로부터 형태소 분석 결과를 가져온다.

5. 분석기의 구현

프로그램은 C 언어를 이용하여 구현하였다. 여기에서 사전은 Clipper에서 제공되는 dbu를 이용하여 구성하였으며, 이 사전에서의 자료의 참조는 C 언어에서 이용할 수 있도록 만들어진 C 라이브러리 함수인 CodeBase를 이용하였다. 형태소 분석을 구현하기 위해 국민학교 6 학년부터 중학교 3 학년까지의 국어 교과서에서 약 27,630 말마디를 표본으로 추출하여 말뭉치를 구성하였다. 말마디 정보와 형태소 접속 특성을 수록한 사전을 이용하여 말뭉치로부터 형태소 분석을 수행한 결과 98 % 이상의 성공율을 보였다.

5.1 분석되는 예

실제로 분석되는 과정을 이해하기 위해 예로서 다음 문장을 살펴보자.

가게에는 여러 가지 물건들을 팔었습니다.

이 문장을 분석하기 위해 첫 말마디 '가게에는'을 입력하여 말마디 화일을 탐색하면, 탐색에 실패함으로써 중의성이 없음을 알 수 있다(단계 1). 따라서 '가게에는'은 접속 특성을 이용한 일반분석기로 처리되어 결과 '가게(명사)+에는(조사)'를 얻는다. 다른 말마디 '여러', '물건들을', '팔었습니다' 등도 중의성을 갖지 않으므로 같은 방법으로서 아

래와 같은 처리 결과를 얻는다.

여러 ---> 여러(관형사)
물건들을 ---> 물건(명사)+들(체언접미사)+을(조사)
팔았습니다 ---> 팔(동사)+았(선어말어미)+습니다(어미)

그러나 여기에서 말마디 '가지'는 (단계 1)과 (단계 2)를 거쳐 형태중의성을 갖는다는 것을 알 수 있다. 따라서 형태중의특성 화일로부터 읽어 온 특성 '11C00'을 보면 앞에 있는 어간형태소에 따라 형태소가 결정됨을 보여 준다(단계 3). 입력문장으로부터 앞에 있는 어간형태소 '여러'를 가져와(단계 4) 형태중의형태 테이블로부터 해당 형태소 결과 인덱스 '100'을 얻는다(단계 5). 인덱스 '100'을 이용해 형태유형 화일로부터 형태소분석 결과인 '가지(00)'의 결과를 가져 오고(단계 6), 품사중의성을 갖지 않으므로 말마디의 형태소분석 결과 '가지 ---> 가지(00)'를 출력한다.

말마디특성

말마디	코드
가지	12

형태중의특성

말마디	색인	특징
가지	1	19000
가지	2	11C00

형태중의형태

말마디	형태소	색인
가지	여러	100

형태유형

말마디	색인	형태1	품사1	형태2	품사2	형태3	품사3	빈도
가지	1	가지	00					82
가지	2	가	06	지	17			18

(그림 6) 말마디 '가지'의 테이블 정보
(Fig. 6) Example Tables of the Word '가지'

6. 분석 결과

중의성 말마디 582개('가지', '갈', '날', '단' 등)의 용례를 300만 말마디의 말뭉치로부터 추출하였으며, 이 문장으로부터 중의성 말마디들의 인접 정보를 분석해 사전에 수록하였다. 말뭉치의 말마디들을 분석한 결과는 <표 2>와 같다. 일반 분석기에서 중의성이 없으나 복수 개의 형태소 분석 결과를 보여주는 말마디(525개)를 완전히 해결하였으며, 일반 분석기에서 전혀 다룰 수 없는 중의성 말마디(582개)의 형태소 분석을 본 분석기는 거의 성공적으로 다루었다.

<표 2> 형태소 분석 결과
<Table 2> Result of Morphological Analysis

말마디	분 류	숫 자	백분율
말뭉치 말마디	총 갯수	4,443	100.00
중의성이 있는 말마디	해결된 말마디	582	13.10
	미해결 말마디	22	0.50
중의성은 없으나 복수결과를 갖던 말마디	해결된 말마디	525	11.80
	미해결 말마디	0	0.00
접속특성에 의해 분석된 말마디	해결된 말마디	3,336	75.10
	미해결 말마디	0	0.00

7. 결 론

지금까지 시도했던 규칙과 접속특성을 이용한 한국어 형태소 분석기에 인접 말마디정보 사전을 보강하여 보다 효율적인 형태소 분석기를 구현하였다. 특히 과거에 전혀 처리하지 못했던 중의성을 갖는 말마디의 분석을 시도하였고, 말마디 사전만으로 형태소 분석을 실시함으로써 야기되었던 공간문제와 탐색시간 문제를 해결하였다. 이 분석기로서 우리는 어휘와 통사적인 정보만을 이용하는 초기 단계의 형태소 분석은 거의 완벽하게 실시될 것이다. 좀더 나은 결과를 얻는 것은 얼마나 많은 중의성 말마디의 인접 정보를 분석해 사전에 수록하느냐에 따라 결정될 것이다. 또한 차후에 의미 정보를 보강함으로써 보다 효율적이며 명확한 형태소 분석 결과를 얻을 수 있으며, 아울러 의미적 중의성을 해결할 수 있을 것으로 기대된다.

참 고 문 헌

[1] L. J. Cahill, "Syllable-based Morphology", Proceedings of the 13th International

Conference on Computational Linguistics (COLING-90), pp.48-53, August 1990.

[2] K. Koskenniemi, "Two-level Model for Morphological Analysis", Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-83), pp.683-685, 1983.

[3] S. Kang, "A Recognition Model of Korean Word Phrase by Syllable Information ", Proceedings of the Korea-US Bilateral Workshop on Computers, Artificial Intelligence and Cognitive Science, pp.292-295, 1991.

[4] 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속정보를 이용한 한국어 형태소 분석기", 춘계 인공지능학술발표회 논문집, pp. 133-147, 1987.

[5] T. Mine, R. Taniguchi, and M. Amamiya, "Coordinated Morphological and Syntactic Analysis of Japanese Language", Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-91), pp.1012-1017, 1991.

[6] G. J. Russel, G. D. Richie, S. G. Pulman and A. W. Black, "A Dictionary and Morphological Analyzer for English", Proceedings of the 11th InternationalConference on Computational Linguistics (COLING-86), pp.277-279, 1986.

[7] 서영훈, "의미 정보를 이용하는 중심어 주도의 한국어 파싱", 서울대학교 공학박사 학위 논문, 1991.

[8] J. Bear, "Morphology and Two-level Rules and Negative Rule Features", Proceedings of the 12th International Conference on Computational Linguistics (COLING-88), pp.28-31, August 1988.

[9] H. Trost, "X2MORF: A Morphological Component Based on Augmented Two-Level Morphology", Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-88), pp. 1024-1030, 1988.

[10] 김덕봉, 최기선, "DDAG : 효율적인 한국어 형태소 해석 방법", 한글 및 한국어 정보 처리 학술발표논문집, pp.341-353, 1993.

[11] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대 대학원 박사학위 논문, 1993.

[12] 남기심, 고영근, 표준 국어 문법론, 탑 출판사, 1985.

[13] 남기심, 이상섭, 김슬옹, 이기황, "한국어 사전의 어휘론적 분석 연구", 우리말 정보화 잔치 학술발표 논문집, pp.327-335, 1991.

[14] 박영환, 윤준태, 송만석, "말뭉치에 근거한 한국어 사전 표제어 구성", 한글 및 한국어 정보 처리 학술발표논문집, pp.58-65, 1991.

[15] 임권묵, "말마디 사전에 의한 미처리 어절과 중의성 어절 분석 연구", 대신대학교 논문집, 제 11 집, pp. 415-427, 1991.

[16] James Allen, Natural Language Understanding, The Benjamin/Cummings Publishing Company, 1987.

[17] 박영환, "말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현", 연세대 대학원 석사학위 논문, 1992.

[18] 신기철, 신용철, 새 우리말 큰 사전, 삼성출판사, 1978.

[19] 임권묵, "형태 중의성 해결을 위한 말마디 사전 설계에 관한 연구 ", 대신대학교 논문집, 제 12집, pp. 341-357, 1992.

임 권 묵

1978년 경희대학교 전자공학과 졸업(학사)

1987년 W. Illinois University Computer Science(이학석사)

1990년 Indiana University Computer Science(박사수료)

1992년~현재 연세대학교 전산과학과 박사과정

1991년~현재 대신대학교 전자계산학과 조교수

관심분야 : 자연어처리(특히, 한국어처리 및 기계번역)

송 만 석

1963년 한남대학교 수학과 졸업 (학사)
1972년 Univ. of Wisconsin 수학과(이학석사)
1978년 Univ. of Michigan 수학과(이학박사)
1981년~현재 연세대학교 전산과학과 교수

관심분야 : 한국어정보처리, 수치해석