

형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기

김 병 희, 임 권 목, 송 만 석
연세대학교 전산과학과

Morphological Analyzer using Adjacent Attribute and Near Word Information

Kim, Byung-hi; Lim, Kwon-mook; Song, Man-suk
Dept. of Computer Science, Yonsei University

요 약

본 논문은 형태소간의 접속 특성과 대형 말뭉치(Corpus)에서 추출된 인접 말마디의 정보를 이용해서 한국어 형태소 분석기를 구현한다. 언어는 단지 규칙으로만 처리하기에는 불가능한 복잡한 구조와 중의성을 갖고 있기 때문에 과거에 주로 연구되었던 형태소들간의 접속 특성과 규칙을 이용한 형태소 분석은 실제로 구문분석 단계에서 사용될 수 있는 실용성을 제시하지 못했다. 따라서 형태소 접속 특성뿐만 아니라 인접 말마디와의 관계를 사전에 기술함으로써 보다 실용성 있는 형태소 분석기의 구현을 시도한다. 아울러 본 형태소 분석기의 효능은 정확하고 풍부한 정보를 사전에 효율적으로 수록함으로써 이득될 것이며, 이를 위해 기존 사전의 보강에 필요한 정보들을 대형 말뭉치로부터 추출하여 사전에 첨가시킨다.

I. 서 론

컴퓨터를 이용한 자연 언어 처리는 컴퓨터 사용자들에게 편리성을 제공하기 위해 최근에 많이 연구되고 있는 분야이다. [1][2] 즉, 컴퓨터가 우리의 일상 언어를 직접 이해하고 생성함으로써 사용자는 더 이상 특수한 프로그래밍 언어나 명령어를 사용하지 않고 컴퓨터와 직접 대화를 할 수 있게 된다. 또한 컴퓨터가 스스로 번역 능력을 갖게 됨으로써 세계 사람들의 언어 장벽을 해소하게 된다.

형태소 분석은 자연 언어 문자열로부터 의미를 추출하기 위한 자연언어 처리 시스템의 첫 단계이다. [3] 자연 언어는 인위적으로 만들어진 형식 언어와는 달리 어휘적, 통사적, 의미적 중의성(ambiguity)이 발생한다. 이러한 중의성을 해결하기 위한 첫 과정으로서, 형태소 분석기는 말마디들을 분석하여 각 말마디가 갖을 수 있는 가능한 모든 형태소 결합 구조를 제시해야 한다. [4]

접속 특성만을 이용했던 기존의 형태소 분석기는 분석된 말마디의 가능한 모든 형태소 결합 구조를 생성함과 동시에 형태소간의 결합 가능 여부를 판단할 수 있는 접속 특성을 이용해 잘못된 결합구조를 제거함으로써 형태소 분석 결과를 제시하였다. [5] 그러나 이러한 정보는 주로 어휘적, 통사적 접속 특성만을 포함할 뿐 의미적인 접속 특성이 결여됨으로써 불

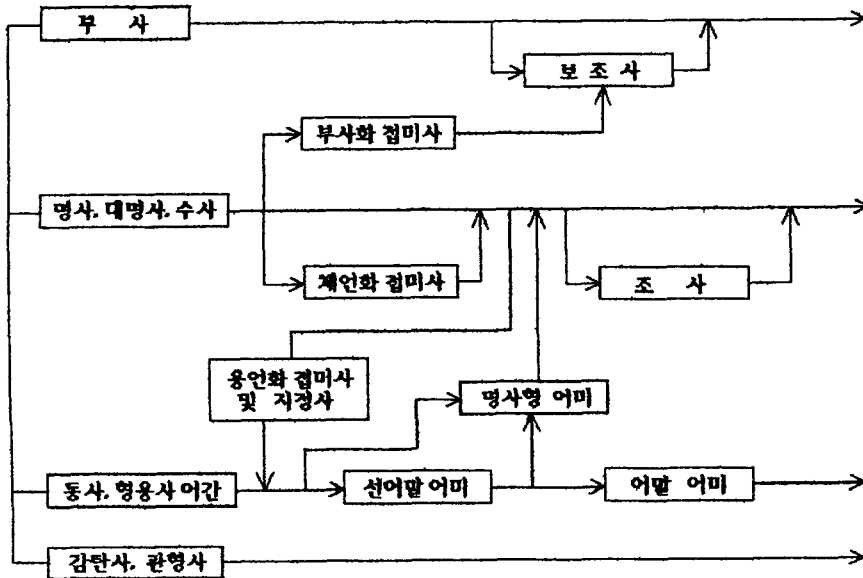
필요한 형태소 결합 구조도 함께 제시하였다. 즉, 어휘적, 통사적 결합 제한은 효과적으로 시도되었으나 의미적인 접속 제한은 적용하지 못함으로써 불필요한 결합 형태가 제시되었고, 이것은 또한 보다 많은 중의성을 야기시켰다.

요즘 새로이 연구된 말마디 사전을 이용한 형태소 분석기는 말마디의 형태소 결합 형태를 일일이 사전에 등록함으로써 형태소 분석 결과를 정확하고 쉽게 제시하였다.[6] 또한 분석 오류로 생기는 중의성도 제거하였고, 중의성을 갖는 말마디도 인접 말마디 정보를 사전에서 참조함으로써 문맥에 맞지 않는 분석 형태를 제거하였다. 그러나 모든 말마디의 분석 형태를 사전에 저장할 수 있는 저장 공간 확보와 말마디의 양이 증가함으로써 소요되는 사전 탐색 시간을 줄여야 한다는 숙제가 남아 있다.[7]

지금까지 앞에서 제시되었던 두 가지 형태의 형태소 분석기를 혼합 구성함으로써 양 분석기가 갖고 있는 단점을 보완할 수 있는 형태소 분석기를 개발한다. 즉, 말마디 사전을 이용함으로써 의미적 접속 특성의 부재로 인해 발생하는 불필요한 결합 형태와 중의성을 효과적으로 제거하고, 접속 특성을 이용함으로써 말마디 사전에 수록해야 할 정보의 양을 현저하게 줄인다.

II. 말마디 구조 분석

하나의 말마디는 하나 이상의 형태소들의 결합으로 구성되어 있으며, 구성 전이도로서 다음과 같이 나타낼 수 있다.[10]



<그림 1> 말마디의 구성 전이도

- (1) 접미사는 체언화 접미사, 용언화 접미사, 부사화 접미사의 세 가지 종류로 분류되며, 각 접미사의 접속 특성에 따라 결합된다.
- (2) 체언화 접미사는 체언 어근의 우측에 접속하여 말마디의 성격을 체언으로 유지하는 성질을 갖고 있다.
- (3) 용언화 접미사는 체언이나 용언 어근의 우측에 접속하여 말마디를 용언의 성격으로 바꾸는 성질을 갖고 있다.

- (4) 부사화 접미사는 체언 어근의 우측에 접속하여 말마디를 부사의 성격으로 바꾸는 성질을 갖고 있으며, '스레', '없이' 등을 말한다.
- (5) 조사는 어근 형태소의 종성에 따라 결합이 달라지며, 보조사는 부사와도 결합이 가능하다. 조사와 조사가 겹쳐서 하나의 조사를 이루는 복합조사는 하나의 조사로 한다.
- (6) 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 결합 형태를 하나의 어미로 처리하였다.
- (7) 명사형 어미는 어말 어미와는 달리 용언 어간의 우측에 결합하여 말마디를 체언의 성격으로 바꿔 주는 역할을 하며 우측에 조사와 결합할 수 있는 특징을 갖고 있다.
- (8) 선어말 어미는 시제, 높임 선어말 어미만을 처리한다. 선어말 어미끼리의 결합은 복합 선어말 어미로 사전에 수록하여 처리하며, '시겠', '시었', '시었겠' 등이 있다.
- (9) 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.

III. 접속 특성을 이용한 분석기

접속 특성을 이용한 형태소 분석은 사전에 형태소와 그 형태소의 좌우 접속 특성을 저장하고, 이 접속 특성을 이용하여 형태소들의 접속 가능 여부를 검사하는 CYK (Cocke-Younger-Kassami) 알고리즘을 이용한다. 지금까지 발표된 대표적인 분석기로는 Tabular Parsing을 이용한 분석법과 역방향으로 분석을 시도한 박영환 분석법이 있다.[11][12] Tabular Parsing 방법은 형태소 접속에 대한 모든 정보를 가지는 좌우 접속 특성표를 작성하기가 매우 어렵고, 사전에 올바른 접속 특성을 부여하기가 쉽지 않으며, 좌우 접속표의 유지 및 관리가 어렵다.[5] 반면에 박영환 분석법은 형태소를 통사적인 특성에 따라 체언, 용언, 조사, 어미로 분류해 접속 특성을 구성함으로써 접속 특성의 저장 및 관리가 용이하다.

3.1 형태소 사전

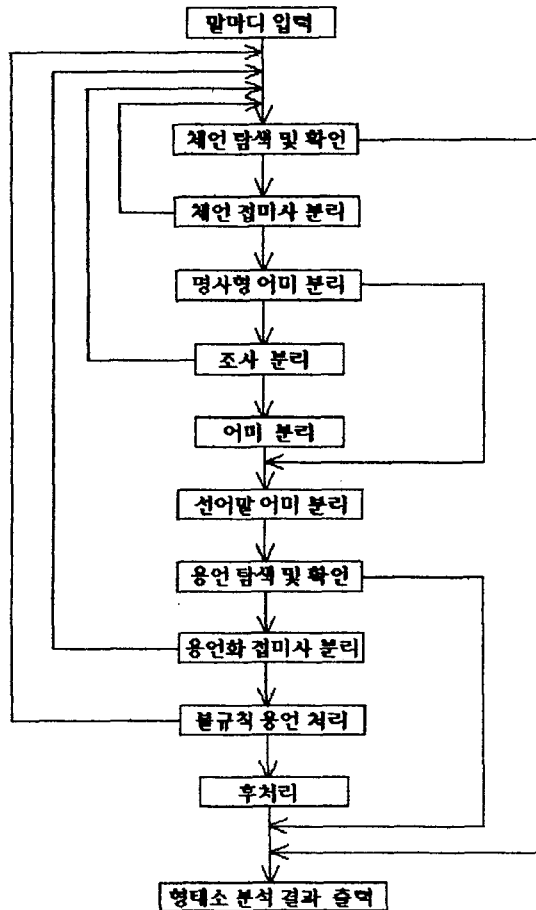
사전은 기능어 사전(function word)과 체언류 사전, 용언류 사전으로 구성된다. 기능어 사전은 조사, 어미, 선어말 어미, 체언화 접미사, 용언화 접미사, 부사화 접미사 사전으로 나누고, 각 사전은 기능어와 각 기능어들의 결합정보를 갖고 있다. 체언 사전은 명사, 대명사, 부사, 감탄사, 수사, 관형사의 품사정보와 명사의 경우 용언화 접미사 '하다', '되다, 와의 결합이 가능한 가를 나타내는 정보를 수록한다. 명사와 명사가 결합해서 이루어진 복합명사는 하나의 명사로서 수록한다. 용언 사전은 동사와 형용사의 어간, 품사정보, 불규칙 정보를 수록한다.[13]

기능어 사전은 분석기에서 자주 사용되고 용량이 작으므로 주기억 장치에 저장 사용한다. 체언 사전과 용언 사전은 보조 기억 장치에 저장하고 필요한 부분만을 분석기 수행 중에 주기억 장치에 적재하여 사용한다.

3.2 분석기의 구현

형태소 분석은 말마디의 우측에서부터 좌측 방향으로 종성자소 혹은 음소 하나씩을 분리하면서 기능어들을 먼저 처리한다. <그림 2>는 분석절차를 전이도로 보여준다.[12]

- (1) 체언 탐색 및 확인 - 입력어(모듈에 입력된 문자열)를 체언사전으로부터 탐색하여 체언 소라는 것을 확인한다. 체언 형태소는 결과를 출력하고 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.



<그림 2> 접속 특성을 이용한 분석 절차

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

- (7) 용언 탐색 및 확인 - 입력어를 용언사전으로부터 탐색하여 용언 형태소라는 것을 확인한다. 용언 형태소는 결과를 출력하고 입력어는 용언 접미사 분리를 시도한다.
- (8) 용언화 접미사분리 - 입력어의 우측부터 음소를 하나씩 분리하여 용언화 접미사임을 확인한다. 용언화 접미사가 제거된 어간은 체언 탐색 및 확인을 시도하고 입력어는 불규칙 용언 처리를 시도한다.
- (9) 불규칙 용언 처리 - 어간 변화형 불규칙 용언('ㄷ', 'ㅅ', 'ㄹ', 'ㅎ')과 어미 변화형 불규칙 용언('어', '거라', '너라', '러', '르', 'ㅂ', '으', '우')을 처리한다.
- (10) 후처리 - 용언의 어간과 어미 사이에 일어나는 음운 축약 현상을 원형 복원하여 처리한다. ('돼' -> '되어', '치' -> '하지' 등)

IV. 말마디 사전을 이용한 분석기

말마디 사전을 이용한 형태소 분석은 말마디 정보와 형태소 분석결과를 사전에 미리 수록하고, 이 말마디 정보를 이용하여 형태소 분석결과를 추출해 낸다. 이 방법은 우리가 이미 알고 있는 말마디의 분석결과를 데이터베이스를 이용해 설계한 사전에 효율적으로 저장함으로써 쉽게 결과를 얻을 수 있다. [6] 또한 규칙으로는 해결할 수 없는 말마디나 중의성을 갖는 말마디는 좌우 인접 말마디의 정보를 참조함으로써 쉽게 분석결과를 얻을 수 있다. 여기서 사용되는 인접 말마디의 정보는 대형 말뭉치를 이용한 용례 분석을 통하여 수집한다. [7]

4.1 말마디 사전 구성

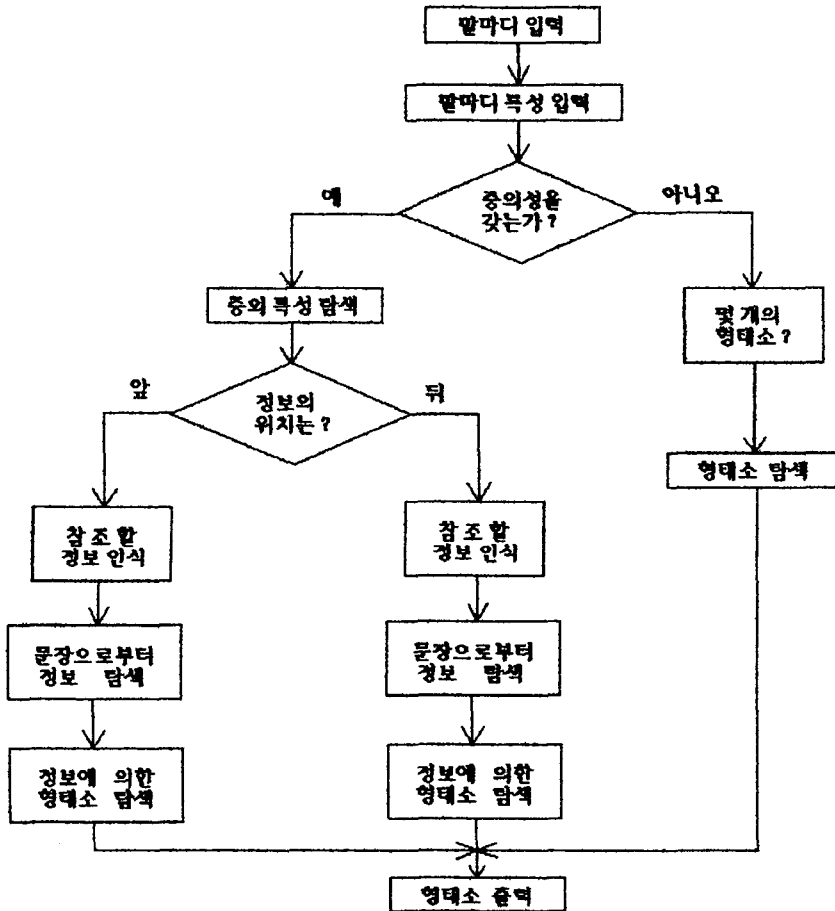
번호	1st	2nd	3rd	4th	5th
1	Front OR	인접어절	인접어절		
2	Front AND	명사형태	명사형태	명사의미	명사의미
3	FR AND	대명사형태	대명사형태	대명사의미	대명사의미
4	RF AND	관형사형태	관형사형태	관형사의미	관형사의미
5	Rear OR	부사형태	부사형태	부사의미	부사의미
6	Rear AND	동사형태	동사형태	동사의미	동사의미
7		형용사형태	형용사형태	형용사의미	형용사의미
8		조사형태	조사형태	조사의미	조사의미
9		어미형태	어미형태	어미의미	어미의미
A		첫형태소형태	첫형태소형태	앞의의미	앞의의미
B		끝형태소형태	끝형태소형태	뒤의의미	뒤의의미
C		첫형태소품사	첫형태소품사		
D		끝형태소품사	끝형태소품사		
E		문두 /미	문두 /미		

< 중의 특성 표 >

말마디 사전은 특성 테이블, 유형 테이블, 인접정보 테이블로 구성된다. 특성 테이블은 각 말마디의 특성과 중의 특성을 코드로서 표현하고, 유형 테이블은 각 말마디가 갖을 수 있는 모든 형태소분석 유형을 제시하며, 인접정보 테이블은 중의성을 해결하기 위한 인접정보와의 관계를 나타낸다. [7]

4.2 분석기의 구현

- (1) 말마디 특성 탐색 - 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.



<그림 3> 말마디 사전을 이용한 분석 절차

- (4) 참조할 정보 인식 - 말마디의 중의 특성을 분석하여 문장으로부터 참조할 정보가 무엇 인가를 인식한다.
- (5) 문장으로부터 정보 탐색 - 참조할 정보를 문장으로부터 탐색한다.
- (6) 정보에 의한 형태소 탐색 - 말마디와 참조한 정보로서 해당 테이블로부터 형태소 분석 결과를 가져온다.

V. 형태소 분석기의 설계

접속 특성과 말마디 사전을 이용한 분석기를 혼합 구성함으로써 말마디의 성격에 따라 분석기를 선택하여 형태소 분석을 실시한다.

5.1 말마디의 분류

말마디의 분석기 실행 기준에 따라 분류한다. [13]

(1) 중의성을 갖는 말마디

한국어가 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. 예를 들면

가지 --> 가지 (명사)
가 (동사) + 지 (어미)

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

단 --> 단 (명사)
단 (부사)

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

달려 --> 달리 (동사) + 어 (어미) : hang
달리 (동사) + 어 (어미) : run

(2) 중의성은 없으나 접속 특성으로 분석이 어려운 말마디

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

얼마 --> 얼 (동사) + 마 (어미)

* 위와 같은 분석 결과는 나타나지 않는 것이 옳으나,
의미 정보의 부재로 인하여 발생하는 분석 결과이다.

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

난 --> 나 (대명사) + 는 (조사)

(3) 접속 특성으로 분석이 용이한 말마디

어휘적 접속 특성과 통사적 접속 특성만으로 형태소 분석이 성공적으로 실행되는 말마디이다.

5.2 사전의 구성

5.2.1 접속 특성을 이용한 형태소 분석을 위한 사전의 구성

5.1에서 (3)과 같은 특성을 가진 말마디로서 체언류 205,701 개(명사 193,341 개, 대명사 211 개, 부사 10,263 개, 감탄사 629 개, 수사 161 개, 관형사 1,096 개)와 용언류 17,642 개(동사 12,027 개, 형용사 5,615 개)를 접속 특성과 함께 입력하였다. 또한 기능어 사전으로는 체언화 접미사 11 개('건', '끼리', '들' 등), 용언화 접미사 9 개('답다', '당하다', '되다' 등), 보조용언 23 개('가다', '만하다', '못하다' 등)를 각각 수록했다.

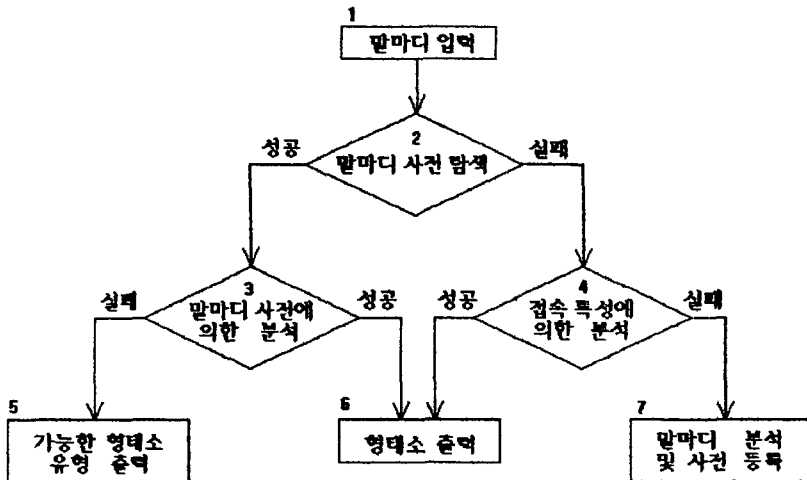
5.2.2. 말마디 사전을 이용한 형태소 분석을 위한 사전의 구성

5.1에서 (1), (2)와 같은 특성을 가진 말마디로서 중의성을 갖는 말마디 582 개('가지', '갈', '날', '단', 등)와 접속 특성으로 분석이 어려운 말마디 525 개('순간', '시간', '알아', '언제나' 등)를 수록했다.

그리고 또한 (1), (2)와 같은 특징을 가지고 있지 않지만 형태소 분석의 효율성을 위하여 특별히 말마디 사전에 입력한 말마디들이 있다. 이들은 대개 접속 특성을 이용하는 경우에 처리하는 데 걸리는 시간이 많이 걸리는 말마디들이다. 예를 들면 말마디의 길이가 다른 말마디들에 비하여 긴 경우이다. 본 논문에서는 5 글자 이상으로 이루어진 말마디들은 말마디 사전에 입력하여 처리 시간의 단축을 꾀하였다. 나중에 사전 표제어의 증가로 인하여 전체적인 분석의 효율이 감소하는 경우에는 말마디 사전의 표제어에서 제거되어야 한다.

5.3 분석기의 설계

- (1) 말마디 입력 - 문장 단위로 말마디들을 분리하여, 말마디별로 분석을 시도한다.
- (2) 말마디 사전 탐색 - 입력 말마디를 말마디 사전으로부터 탐색하여 등록되어 있는가를 확인한다.
- (3) 말마디 사전에 의한 분석 - 말마디 사전을 이용한 분석 방법으로서 말마디 사전에 등록되어 있는 말마디의 형태소 분석을 실시한다.



<그림 4> 형태소 분석 절차

- (4) 접속 특성에 의한 분석 - 접속 특성을 이용한 분석 방법으로서 말마디 사전에 등록되지 않은 말마디의 형태소 분석을 실시한다.

- (5) 가능한 형태소 유형 출력 - 말마디 사전에 등록되어 있으나 분석에 실패한 말마디의 가능한 모든 유형을 제시한다.
- (6) 말마디 사전이나 접속 특성을 이용하여 분석에 성공한 결과를 출력한다.
- (7) 말마디분석 및 사전등록 - 말마디 사전에 등록되어 있지 않으며 분석에 실패한 말마디를 분석하여 말마디 사전에 등록한다.

5.4 분석의 예

- (1) 말마디 사전의 정보를 이용하여 성공적으로 분석되는 경우

우리 나라 사람은 서로 사랑하며 잘 삽니다.

위의 문장에서 '삽니다'는 말마디 사전의 정보를 이용하여(바로 앞 어절의 어미의 형태 '며'의 정보를 이용) <그림 4>에서 동사 '살-'과 어미 'ㅂ니다'로 ①→②→③→⑥의 과정을 거쳐 분석되어진다.

- (2) 말마디 사전의 정보를 이용하여 부분적으로 성공되어지는 경우

내가 세상에서 가장 귀하게 생각하는 단 하나는 부모님 뿐이다.
강단의 단 하나가 급이 생긴 상태이다.

위의 문장에서 '단'은 말마디 사전의 정보를 이용하더라도 증의성이 없이 하나의 결과로만 분석이 되지를 않는다. 따라서 위의 문장에서 말마디 '단'은 <그림 4>에서 ①→②→③→⑤의 과정을 거쳐 가능한 결과를 최대한으로 줄여 분석이 가능한 명사 '단'과 관형사 '단'의 두가지 가능한 분석 결과가 출력되어진다.

- (3) 접속 특성을 이용하여 분석이 성공되는 경우

가을 하늘은 유난히 파랗다

위의 문장에서 '하늘은'과 같은 어절은 접속 특성을 이용하여 명사 '하늘'과 조사 '은'으로 <그림 4>에서 ①→②→④→⑥의 과정을 거쳐 분석되어진다.

- (4) 접속 특성을 이용하여 분석이 실패하는 경우

이종범의 환상적인 수비로 해태는 7 회의 위기를 넘기고 승리했다.

위의 문장에서 '이종범의'와 같은 어절은 접속 특성을 이용한다 하더라도 체언 사전에 '이종범'이라는 표제어가 없기 때문에 성공적으로 분석할 수가 없다. 따라서 '이종범의'이라는 말마디는 <그림 4>에서 ①→②→④→⑦의 과정을 거치게 되므로 이를 성공적으로 분석할 수 있게 하기 위해서는 체언 사전에 '이종범'이라는 체언을 등록하여야만 하게 된다.

5.5 분석기의 구현

프로그램은 C를 이용하여 구현하였다. 여기에서 사전은 Clipper에서 제공되는 dbu를 이용하여 구성하였으며, 이 사전에서의 자료의 참조는 C에서 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 발음치를 구성하였다. 말마디 정보와 형태소 접속 특성을 수록한 사전을 이용하여 발음치로부터 형태소 분석을 수행한 결과 95 % 이상의 성공율을 보였다.

VI. 결론

지금까지 시도했던 규칙과 인접특성을 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 인접정보를 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.

VII. 참고문헌

- [1] James Allen, Natural Language Understanding, The Benjamin/Cummings Publishing Company, 1987.
- [2] Gerard Salton, Automatic Text Processing, Addison Wesley, 1989.
- [3] 김대식, "Lexical Database 구축을 위한 어절분석도구에 관한 연구", 연세대 대학원 석사학위 논문, 1991.
- [4] 양단희, "한국어 전자사전 원형의 설계 및 구현", 연세대 대학원 석사 학위 논문, 1991.
- [5] 강승식, 김영택, "사전 정보에 기반한 효율적인 한국어 형태소 분석기의 설계 및 구현", 춘계 정보과학회 학술발표논문집, 1991.
- [6] 임권득, "어절 사전에 의한 미처리 어절과 중의성 어절 분석 연구", 대신대학 논문집, 제 11 집 pp. 415-427, 1991.
- [7] 임권득, "형태 중의성 해결을 위한 말마디 사전 설계에 관한 연구", 대신대학교 논문집, 제 12 집 pp. 341-357, 1992.
- [8] 남기심, 표준 국어 문법론, 탑 출판사, 1985.
- [9] 유찬호, "한국어 문장의 개념 분석에 관한 연구", 동국대학교 석사학위 논문, 1987.
- [10] 남기심, "한국어 사전의 어휘론적 분석 연구", 우리말 정보화 잔치 학술발표 논문집, 1991.
- [11] 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속정보를 이용한 한국어 형태소 분석기", 춘계 인공지능학술발표회 논문집, 1987.
- [12] 박영환, "말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현", 연세대 대학원 석사학위 논문, 1992.
- [13] 박영환 외 2인, "말뭉치에 근거한 한국어 사전 표제어 구성", 한글 및 한국어 정보 처리 학술발표논문집, 1991.'