음절에 기반한 한국어 형태소 분석기

장 동수, 서 영훈
충북대학교 컴퓨터공학과

Syllable-Based Korean Morphological Analyzer

Dong-Su Jang, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

본 논문에서는 한국어의 음절 특성을 이용한 한국어 형태소 분석기를 제시하였다. 이 형태소 분석기는 품사별 음절 정보, 불규칙 음절 정보, 활용어절 음절 정보, 선어말 어미 음절 정보 등을 이용하여 음절 단위로 형태소 분석을 한다. 음절 단위의 형태소 분석 방법은 음소 단위의 방법보다 형태소 분석시에 생성될 수 있는 잘못된 중간 분석 결과를 크게 감소시켜, 사전 탐색 부담을 최소화한다.

시스템의 사전은 품사별 결합 특성과 사전 표제어의 길이별 분포 특성을 이용하여 구성하였으며, 그 규모는 약 16만 어휘이다. 이러한 사전 구성은 효율적인 사전검색을 제공하며, 특히 철자 검색기와 자동 인덱싱 등의 다양한 응용 시스템 요구를 곧바로 수용할 수 있는 유연성과 효율성을 갖고 있다.

I. 서 론

자연언어 처리는 사용자에게 편리한 인터페이스를 제공하고, 정보화 시대에서 급증하는 정보교류를 원활히 하기 위해 필수적인 연구 분야이며, 이미 선진 각국에서 차세대 컴퓨터

시스템 개발을 위해 연구중인 중요 연구분야중 하나이다.

자연 언어 처리 시스템들은 모두 언어학적 지식과 전산학을 결합한 언어분석에 기반을 두고 있으며, 자연 언어 분석은 크게 형태소 분석, 구문 분석, 의미 분석으로 구분되고 있다. 이중 형태소 분석은 통사 분석과 의미 분석의 전단계로서 기계 번역이나 자연 언어 인터페이스, 정보 검색 등 모든 자연언어 관련 분야에 필수적이다. 특히 한국어, 일어처럼 문법형태소의 기능에 의해 단어의 통사적, 의미적 역할이 결정되는 교착어에서는 형태소 분석이 통사 분석과 의미 분석에 미치는 영향이 크기 때문에 한국어의 분석에 있어서 형태소 분석이 아주 중요하다고 할 수 있다. 그러나, 지금까지 한국어의 형태소 분석에 관심이 적었던 이유는 영어, 불어, 독일어 같은 굴절어의 형태소 분석이 한국어에 비해 매우 간단해서 형태소 분석을 당연한 것으로 간주했기 때문이다.

기존의 시스템들이 주로 사용하고 있는 형태소 분석방법은 형태론적 변형 현상을 처리하는 방법 즉, 근접한 형태소 간의 결합조건을 검사하여 형태소의 일부를 교체(replacement), 삽입(insertion), 삭제(deletion)하는 방법이다. 그러나, 하나 이상의 형태소가 결합하여 한 단어를 이루고 있는 교착어에서는 형태론적 변형 현상의 처리보다는 형태소의 분리가 선행되어야 한다. 그러므로 한국어의 형태소 분석은 단어를 이루는 형태소를 분리한 후에, 형태론적 변형이 일어난 형태소의 원형을 복원하고, 사전과 단어 사이의 결합관계로 옳은 분석 결과를 찾아내는 과정이라 할 수 있다.

전산 언어학에서 형태론은 모든 언어에 공통되는 현상을 처리하기 위한 형태소 분석론과 특정 언어의 언어 현상을 처리하기 위한 형태소 분석론으로 나눌 수 있다. 전자에는 two-level morphology[Kosk 83]와 syllable-based morphology[Cahi 90]가 있다. 그리고, 한국어와 관련된 후자의 형태론에는 단어를 검색하는 방향에 따라 문자열을 왼쪽에서 오른쪽으로 검색하는 좌우 분석법, 오른쪽에서 왼쪽으로 검색하는 우좌 검색법, 양쪽에서 동시에 검색하는 양방향 검색법이 있고[박종만 90], 단어를 이루는 형태소의 길이에 따라 단어를 가능한 모든 형태소로 분할한 다음 그 단어를 이루고 있는 형태소들의 집합 중 가장 긴 형태소를 포함하고 있는 결과를 선택하는 최장 일치법(longest match strategy), 가장 짧은 형태소들로 이루어진 집합을 선택하는 최단 일치법(shortest match strategy)[김덕봉 90], head-tail 구분법[최형석 84], tabular parsing method[김성용 87] 등이 있다.

한글의 음절 특성 정보를 사용하지 않은 기존의 한국어 형태소 분석 시스템들은 문법 형태소를 처리하는 방법과 규칙을 기술하는 방법, 그리고 단어에서 각 형태소를 분리하는 방법에 한계가 있다. 한국어 형태소 분석 시스템에 two-level 방법을 적용하면, 형태소 분석시에 많은 규칙이 적용되어 불필요한 중간 분석 후보가 과다하게 생성되고, 이로 인하여 사전 탐색의 부담이 커지며 효율성이 감소한다. 한국어 형태소 분석 시스템에 최장 일치법,

최단 일치법, 양방향 분석법을 적용하면 문자열의 어느 부분(앞부분, 뒷부분)부터 적용하느냐에 따라 분석 결과가 다르게 나타나며, head-tail 구분법과 tabular parsing mothod는 모든 형태소들의 접속관계를 나타내는 접속 정보표를 구성하기 어려워서 처리 범위가 좁아지는 문제점이 있다.

본 논문의 한국어 형태소 분석 시스템은 음절 특성 정보를 사용하여 음절 단위로 각 형태소를 분리한다. 한글의 음절 특성 정보는 사전의 표제어로 사용되는 음절을 문법 형태소에서 사용되는 음절과 어휘 형태소에서 사용되는 음절들로 구분하여 나타낸 것으로, 이 정보는 형태소를 분리할 때 사용된다. 음절 특성 정보를 이용한 음절 단위의 형태소 분석은, 각 형태소의 분리 위치를 미리 추정하고, 분석될 가능성이 없는 형태소 분리를 미리 제거하여 시스템이 처리해야 하는 중간 결과를 최소화한다. 이러한 특성때문에 음절 단위의 형태소 분석 방법은 기존의 음소 단위 분석 방법으로 한국어를 분석할 때 발생하는 과다한 중간 분석 결과의 수를 크게 감소시켜 사전 검색의 효율성을 증가시킨다.

II. 음절 정보를 이용한 형태소 분석

일반적으로 자연언어를 분석하는 단위는 자소 단위이다. 그러나 자음과 모음이 결합하여 음절을 이루고 다시 하나 이상의 음절이 모여 하나의 단어를 이루는 한글의 글자체계 특성으로 인하여 한국어 분석시에는 자모 단위의 처리와 음절 단위의 처리가 모두 가능하다.

초성, 중성, 종성으로 구성되는 가능한 모든 음절의 수는 초성이 자음 19개, 중성이 모음 21개, 종성이 자음 27개로 구성되므로, 총 11,172(19*21*27)개이다[변정용 92]. 그런데 실제로 현대 국어에서 사용되는 음절의 수는 약 2,350개로 전체의 1/4밖에 되지 않는다. 그리고 형태소 분석기에서는 훨씬 더 적은 1,999개의 음절을 사용한다[강승식 93].

만약에, 어떤 음절이 용언의 표층형으로만 사용된다면, 그 음절이 나타나는 단어에서 그 음절의 앞부분은 어간이고 뒷 부분은 어미라는 사실을 쉽게 알 수 있다. 또한 문법 형태소와 결합하는 어휘 형태소의 끝음절의 특성을 이용하면 문법 형태소를 분리할 때 분리되지 않는 위치를 좀더 명확히 알 수 있다. 이러한 음절 특성 정보를 형태소 분석시에 사용하면 매우 유용하다[강승식 93].

2.1 문법 형태소

음절 단위로 조사를 분석하기 위해서는 조사의 첫음절로 사용되는 음절 정보와 조사의 두번째 음절부터 끝음절까지 사용되는 음절 정보가 필요하다. 이 음절 정보는 어절에서부터 조사가 분리되는 위치를 추정하기 위하여 사용된다.

음절(S_1, S_2, S_3, …, S_n)으로 이루어진 단어에서 음절 S_i가 조사의 첫음절로 사용되고 S_{i+1}, S_{i+2}, …, S_n이 조사의 두번째 이상의 음절로 사용된다면 조사 사전을 찾지 않고도 음절 S_i가 조사의 시작점이라는 것을 쉽게 추정할 수 있다[강승식 93].

음절 단위로 어말 어미를 분석하기 위해서는 어미의 첫음절로 사용되는 음절 정보와 어미의 두번째 음절부터 끝음절까지 사용되는 음절 정보가 필요하다. 이 음절 정보는 어절에서부터 어미가 분리되는 위치를 추정하기 위하여 사용된다.

음절(S_1, S_2, S_3, …, S_n)으로 이루어진 단어에서 음절 S_i가 어미의 첫음절로 사용되고 S_{i+1}, S_{i+2}, …, S_n이 어미의 두번째 이상의 음절로 사용된다면 어미 사전을 찾지 않고도 음절 S_i가 어미의 시작점이라는 것을 쉽게 추정할 수 있다[강승식 93]. 어말어미는 조사와는 달리 불규칙 현상으로 인하여 형태변이가 일어나므로, 형태변이가 일어난 어미에 대하여 따로 분석할 수 있어야 한다.

음절 단위로 선어말 어미를 분석할 때는 단어를 이루고 있는 음절 중에 받침으로 ' 从' 이 있거나 음절 '시' 를 포함하는 단어를 선어말 어미가 포함된 단어로 생각하고, 매개모음과 선어말 어미 변이체 및 축약형을 포함한 결합관계와 결합 조건을 검사하면 선어말 어미가 분리되는 위치를 추정할 수 있다.

2.2 어휘 형태소

음절 정보를 이용한 어휘 형태소 분석은 크게 규칙 형태소와 불규칙 형태소 2가지로 나누어 분석을 한다.

규칙 형태소(체언+조사, 규칙 용언+어미, 부사+조사)는 형태소의 품사/길이별 음절 정보를 이용하면 분리가 매우 간단하다.

음절(S_1, S_2, S_3, …, S_n)으로 이루어진 단어에서 음절 S_i가 문법 형태소의 첫음절로 사용되고 S_{i+1}, S_{i+2}, …, S_n이 문법 형태소의 두번째 이상의 음절로 사용되며 음절 S_{i-1}이 어휘 형태소의 끝음절로 사용된다면 음절 S_i가 어휘 형태소와 문법 형태소의 분리 시작 음절이라는 것을 알 수 있다.

불규칙 형태소는 불규칙 용언에서만 나타나며, 이러한 불규칙 현상은 특정한 음절로 끝나는 용언에서만 발생한다. 그러므로 이러한 각 불규칙 용언의 끝음절을 불규칙 형태소의 처리시에 이용하면, 단어의 불규칙 추정이 아주 쉽게 된다.

음절 (S_1, S_2, S_3, …, S_n)으로 이루어진 단어에서 음절 S_i가 불규칙 용언의 끝음절로 사용되는 음절이라면, 음절 S_i에 대해 원형 복원을 한 후, 복원된 음절 S_i가 어휘 형태소의 끝음절로 사용되는지를 검사한다. 그리고 끝으로 음절 S_{i+1}이 문법 형태소의 첫음절로 사용되고 S_{i+2}, …, S_n이 문법 형태소의 두번째 이상의 음절로 사용된다면 단어의 음절 S_i에서 불규칙 현상이 발생하였다는 사실을 알 수 있다.

한국어에서는 한 어절이 부사, 관형사, 감탄사, 명사, 대명사 등의 독립 형태소로 구성될 수 있으며, 이러한 독립 형태소는 형태소의 품사/길이별 음절 정보를 이용하면 추정이 매우 간단하다.

음절 (S_1, S_2, S_3, …, S_n)으로 이루어진 단어에서 음절 S_n이 어휘 형태소의 끝음절로 사용된다면, 이 단어는 독립 형태소로 존재 가능하다는 사실을 알 수 있다.

III. 사전 구성

기존의 시스템은 한국어 형태소 분석을 위해서 국어사전의 전체 어휘를 문법 형태소와 어휘 형태소로 분리하여 구성하고 있으나, 본 논문에서는 사전 탐색시의 부담을 최소화하기 위해 약 16만 어휘의 사전[금성출 92]을 품사와 길이별로 분류하였다. 시스템의 사전은 품사별로는 각 형태소의 결합 여부와 독립성에 따라 체언(명사, 대명사, 수사), 용언(동사, 형용사), 부사, 독립어(관형사, 감탄사), 문법 형태소(조사, 어미) 사전 등으로 나누었다. 사전의 길이별로는 각 품사의 길이 분포를 고려하여 사전 탐색시에 부담이 큰 체언, 용언, 부사 사전의 각 표제어가 1/3씩 분리되도록 나누어서, 체언 및 용언 사전은 "2음절 이하, 3음절, 4음절 이상" 의 3가지로, 부사 사전은 "2음절 이하, 3음절 이상 4음절 이하, 5음절 이상" 의 3가지로 분류하였다. 어휘 형태소 사전을 체언과 용언 사전으로 분리한 이유는 조사 앞에는 체언과 부사만이 오고, 어미 앞에는 용언만이 오는 결합 특성을 이용하기 위해서 분리하였고, 체언, 용언, 부사 사전을 길이별로 분리한 이유는 사전의 탐색 부담을 줄이고자 분리한 것이다.

사전을 탐색하기 위하여 각 형태소의 길이와 품사 정보를 이용하므로 형태소와 길이, 품

사가 일치하지 않는 사전 표제어는 전혀 검색하지 않게 된다. 이와같은 특징은 자동 인덱싱이나 철자 검색기 등의 응용 시스템에서 단지 필요한 일부분의 사전만 탐색하도록 하는 유연성을 제공한다.

IV. 시스템 구성

대부분의 형태소 분석 시스템은 시스템 내부의 한글 코드로 3-BYTE나 N-BYTE 코드를 사용해 왔다. 그러나, 이 코드 체계들은 자소 단위의 처리는 간단하지만, 음절 단위의 처리를 위해서는 한글 오토마타를 구성해야 하는 어려움이 있다.

본 논문에서는 자소 단위 처리와 음절 단위 처리가 모두 가능하도록 하기 위하여 2-BYTE 상용조합형 코드를 사용하고 있으며, 이 코드는 기존의 다른 코드에 비해 사전 구성에 있어서도 메모리의 절약에 효율성을 제시해 준다.

(그림 1)은 형태소 분석 시스템의 구조를 설명한 것이고, (그림 2)는 단어에 사용하는 문자를 나타낸 것이다.

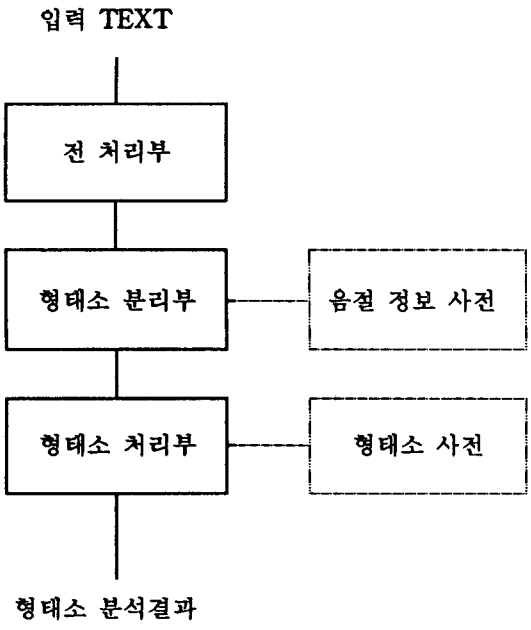

그림 1. 음절 정보에 기반한 한국어 형태소 분석기 구조

문장부호	, . ? ! ‘ ’ “ ” ( ) [ ] { } ; : $ + - * / … 〈 〉
숫 자	1, 2, 3, 4, 5, 6, 7, 8, 9, 0
영문자	A - Z , a - z

그림 2. 단어에서 쓰이는 문장 부호, 영문자, 숫자

4.1 전처리부

(1) 상용 조합형 한글 문서를 입력으로 받아들여 그것을 단일 문장 즉 종지부(‘.’ , ‘?’ , ‘!’ 등)가 어절의 끝에 나타날 때까지를 기준 처리 단위로 하고, 이것을 String으로 저장한다.
(2) 기준 처리 단위인 단일 문장을 다시 띄어쓰기 단위인 어절로 묶어서 어절별로 저장한다. 좌괄호 문장부호(“, (, ‘, 〈, …)가 있으면, 우괄호 문장부호(” ,), ’ , 〉, …)가 있을 때까지 공백(띄어쓰기)을 무시하고, 한 어절에 속하는 음절로서 저장한다. 영문자나 숫자에서도, 영숫자 사이에 공백이 있어도 이것을 무시하고 한 어절에 속하는 음절로 저장한다. 이렇게 띄어쓰기 단위로 구분된 음절들을 실제 분석 단위로 사용한다.
(3) 실제 분석 단위인 단위 어절을 음절별로 각각 저장하고, 단위 어절의 각 음절을 처리할 음절과 처리하지 않을 음절(문장부호)로 나누어 저장한다.

4.2 형태소 분리부 : 음절 정보를 이용하여 좌에서 우로 어절을 분리한다.

(1) 음절이 문장부호로 시작되면, “문장부호 + 조사” , “문장부호 + 접미사 + 어미/조사” , “문장부호 + 어미” 등의 형태일 수 있으므로, 우선 좌괄호 문장부호에서부터 우괄호 문장부호까지를 단일 체언(명사)으로 보고 각 구성 형태에 대하여 음절 정보를 이용하여 분리를 시도한다.
(2) 음절이 문장부호로 시작되지 않으면, 음절 정보를 이용하여 처리할 음절의 끝음절이 독립어(부사, 관형사, 감탄사, 대명사, 명사, …) 음절로 사용되는지를 검사하여, 독립 형태소를 찾아낸다.
(3) 처리할 음절이 2음절 이상일 때, 문법 형태소를 중심으로 각 형태소간의 분리점을 찾아, 어절을 “체언+조사, 용언+어미, 체언+접미사+조사, 용언+접미사+어미” 형태로 분리한다.

(4) 처리할 어절이 불규칙 변형 음절이나 선어말 어미 음절을 포함하고 있으면, 해당 음절의 원형을 복원하고 복원된 원형 음절에 대하여 (3)번의 처리를 수행한다.
(5) 음절정보는 조사, 어미, 체언(2음절, 3음절, …), 용언, 부사, …등의 첫음절 및 끝음절로 사용되는 음절 데이타를 말한다. 그리고 분석부에서 분리되고 저장된 결과는 중간결과로서 사전검색을 거쳐야 한다.

4.3 형태소 처리부 : 사전정보를 이용하여 처리

(1) 중간 분석 결과의 길이와 품사 정보를 가지고 사전의 시작위치를 선택하여 사전을 검색하고, 분리된 형태소들 중 어느한 형태소라도 없으면, 해당 형태소를 포함하고 있는 분석 결과의 나머지 형태소들을 검색하지 않도록 하며, 최종적으로는 남아 있는 결과만을 옳은 분석 결과로서 출력한다.
(2) 중간 분석 결과 중에 처리되지 않은 결과에 대하여, 다시 복합 명사 추정 및 미등록어 추정을 하여, 옳바른 분석결과가 틀린 분석으로 오인되지 않도록 한다.

V. 결론

본 논문에서는 한국어의 음절 특성 정보와 품사별 사전을 이용한 한국어 형태소 분석기를 제시하였다. 형태소를 분석할 때 음절 특성 정보를 사용함으로서, 형태소간의 분리 위치와 미등록어 및 복합어를 미리 추정하였다. 이것은 분석 가능성이 없는 중간 분석 후보를 사전 탐색이나 결합 관계 검사를 거치지 않고 미리 제거할 수 있게 한다. 그리고, 조사와 어미를 통합형으로 처리하여 문법 형태소간의 결합관계를 고려하지 않아도 되도록 구현하였다. 특히 불규칙 현상에 음절 특성 정보를 이용함으로서, 분석 가능성이 없는 불규칙 현상에 대한 처리를 미리 제거하여, 불규칙 형태소의 분석 과정에서 생성될 수 있는 과다한 분석 후보의 수를 줄였다. 사전의 품사/길이별 분류는 시스템의 효율성에 가장 큰 영향을 미치는 사전 탐색의 부담을 줄였으며, 자동 인덱싱이나 철자 검색기 같은 응용 시스템의 제한적 사전 요구에도 곧바로 효율적으로 사용할 수 있는 유연성을 제시하였다.

앞으로의 연구과제로는 서로 연관되어 나타나는 보조 용언들의 통합 처리 문제와 전문 용어 및 신조어 등의 신속한 사전 입력 문제, 그리고 인명처럼 둘 이상의 어절이 하나의 의미를 이루는 어절 구조에 대한 처리 문제가 있다. 또한 사전 검색시에 옳바른 위치로 가기 위해 사용하는 사전 인덱스 값을 사전 갱신시에 자동적으로 처리하는 방법에 대한 연구도 요

구된다. 그리고, 음절 단위 형태소 분석시에 가장 기본적으로 요구되는 한글 음절 특성 정보의 계속적인 수집, 확장도 필요하다.

참고문헌

[강승식 91] 강 승식, 김 영택, "한국어 형태소 분석기에서 선어말 어미의 분석 모형", 한국어정보과학회 논문지, 18권, 5호, pp.505-513, 1991.

[강승식 92] 강 승식, 김 영택, "한국어 형태소 분석기에서 불규칙 용언의 분석 모형", 한국어정보과학회 논문지, 19권, 2호, pp.151-164, 1992.

[강승식 93] 강 승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울 대학교 공학박사 학위논문, 1993.

[김덕봉 90] 김 덕봉, "한국어 형태소 처리와 사전-접속정보를 이용한 한글 철자 및 띄어쓰기 검사기", 어학연구, 26권, 1호, pp.87-113, 1990.

[김성용 87] 김 성용, "Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기", 한국과학기술원, 석사학위논문, 1987.

[금성출 92] 금성출판사 사서부, **뉴에이스 국어사전**, 금성출판사, 1992.

[박갑수 93] 박 갑수, 조 규빈, **하이라이트 고교문법**, 지학사, 1993.

[박종만 90] 박 종만, "효율적인 한국어 형태소 분석기 및 철자 검사 교정기의 구현", 서울대학교 공학석사 학위논문, 1990.

[변정용 92] 변 정용, "훈민정음 창제원리와 한글코드 제정원리 : 자소형 제안", 동국대학교 전자계산학과, 1992.

[최형석 84] 최 형석, "자연어 처리 앨고리즘", 한국어정보과학회 추계 학술발표회 논문집, 11권, 2호, 1984.

[Cahi 90] Cahill, L.J., "Syllable-Based Morphology", Proceedings of the 13th International Conference on Computation Linguistics, vol3, pp.48-53, 1990.

[Kosk 83] Koskenniem, K., "Two-Level Model for Morphological Analysis", Proceedings of the 8th International Joint Conference on Artificial Intelligence, pp.683-685, 1983.