MADE : 형태소 분석기 개발 환경☆

MADE : Morphological Analyzer Development Environment

심 광 섭*
Kwangseob Shim

요 약

본 논문은 실용적인 한국어 형태소 분석기 개발에 사용될 수 있는 도구인 MADE를 소개한다. MADE는 형태소 사전에서 제공되는 인접 조건만을 사용하여 형태소 분석을 수행한다. 이것은 형태소 분석기를 개발하기 위해 별도의 프로그래밍은 전혀 하지 않고 단지 형태소 사전만 구축하면 된다는 것을 의미한다. MADE는 형태소 사전을 구축하고 검증하는 데 필요한 기능들을 제공한다. 일단 형태소 사전이 구축되고 나면 MADE는 독립된 형태소 분석기로서 사용될 수도 있고 형태소 분석기를 필요로 하는 다른 응용 소프트웨어에 내장되어 사용될 수도 있다.

Abstract

This paper proposes a software tool MADE that is useful to develop a practical Korean morphological analyzer. A morphological analysis is performed by using adjacency conditions provided by a morphological dictionary. This means that developing a morphological analyzer is reduced merely to constructing a morphological dictionary. No programming skill is required in this process. MADE provides with useful functions that facilitate the construction of a dictionary. Once a dictionary is constructed, the morphological analysis engine embedded in MADE may be used as a stand-alone morphological analyzer or be integrated into an application software which requires a Korean morphological analysis module.

Keyword : 형태소 분석 (Morphological Analysis), 자동 색인 (Automatic Indexing), 개발 도구 (Developing Tool), 2 단계 형태소 분석 모델 (Two-Level Morphological Analysis Model)

1. 서론

현재까지 국내에서 개발된 한국어 형태소 분석기는 패키지 형태로 보급되었다. 이러한 형태소 분석기에서는 사용자가 임의로 품사 체계를 변경하거나 형태소 분석 결과를 변경할 수 없었다. 또 형태소 분석기 개발 단계에서 분석 오류를 제거하기 위한 노력은 많이 하겠지만 가능한 모든 어절에 대한 전면적인 검증을 하는 것은 아니므로 개발자가 미처 발견하지 못한 오류들이 많을 수밖에 없다. 그런데 형태소 분석기가 패키지 형태로 보급되기 때문에 사용 중에 분석 오류가 발견되더라도 사용자가 직접 오류를 수정할 수 있는 방법은 없었고, 개발자에게 보고하여 수정을 의뢰하는 수밖에 없었다.

국내에서는 패키지 형태의 한국어 형태소 분석기만 개발된 것과 달리 외국에서는 사용자가 직접 형태소 분석기를 만들어 볼 수 있는 도구들이 많이 개발되었다. [1]에서는 이러한 도구에 대하여 폭넓게 소개가 되어 있다. 이들 도구 중에는 형태소 분석 과정을 잘 이해할 수 있도록 학생들의 실습용으로 개발된 것들도 있고 연구용 형태소 분석기를 만들기 위한 도구도 있다. 여러 언어에 적용할 수 있는 도구도 몇몇 있으나 대부분은 영어나 독일어 등과 같은 특정 언어에 대해서만 적용할 수 있는 것들이다. 물론 유럽어를 대상으로 개발된 도구라 하더라도 한국어나 핀란드어와 같이 유럽어에 비하여 형태론적으로 복잡한 언어

* 정 회 원 : 성신여자대학교 컴퓨터정보학부 교수
shim@sungshin.ac.kr
[2007/06/12 투고 - 2007/06/19 심사 - 2007/07/02 심사완료]
☆ 이 논문은 2005년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

에 대해서도 적용할 수 있는 가능성은 있다. 하지만 이는 어디까지나 가능성을 확인해 보는 수준이며 한국어를 대상으로 개발된 도구와 비교하면 성능이나 개발 편의성이 떨어질 수밖에 없을 것이다.

여러 도구 중에서 핀란드의 Koskenniemi가 제안한 Two-Level 형태소 분석 모델에 따라 형태소 분석을 할 수 있는 PC-KIMMO란 도구가 널리 알려져 있다[2, 3]. Two-Level 형태소 분석 모델은 표층형(surface form)에서 어휘형(lexical form)에 이르는 중간 단계의 유도 과정은 생략하고 Two-Level 규칙으로 표층형에서 어휘형으로 직접 변환함을 전제로 한다. 이 모델은 핀란드어를 비롯한 여러 언어에 적용된 바 있으나 한국어 형태소 분석에는 부적합한 모델이라는 지적을 받았다[4]. 일각에서는 이 모델이 한국어 형태소 분석에 적용될 수도 있음을 보여주는 연구 결과도 있다[5, 6]. 하지만 이러한 시도는 어디까지나 Two- Level 형태소 분석 모델이 한국어 형태소 분석에 적용될 수도 있음을 보여주는 초보적인 연구 결과일 뿐, 이 모델로 실제로 응용이 가능한 본격적인 한국어 형태소 분석기를 개발할 수 있음을 입증한 것은 아니다. 일반적으로 실제 응용에 사용되는 한국어 형태소 분석기의 정확도가 99%를 상회하는 수준임에 비하여 [5]에서는 전혀 성능에 대한 보고가 없고 [6]에서는 Two-Level 모델로 구현한 한국어 형태소 분석기의 정확도는 95%라고 보고되었다.

본 논문에서는 우리 국어의 형태론에 대한 지식과 형태소 분석 과정에 대한 간단한 이해만 있으면 별도의 프로그래밍 과정을 거치지 않더라도 실제 응용이 가능한 수준의 한국어 형태소 분석기를 개발할 수 있는 도구인 MADE(Morphological Analyzer Development Environment)를 소개한다. MADE는 형태소 사전에서 제공되는 인접 조건 검사만으로 형태소 분석을 수행하기 때문에 형태소 분석에 필요한 정보가 분석기 내부에 전혀 프로그래밍되어 있지 않으며, 형태소 사전만 만들면 형태소 분석기를 개발할 수 있다.

MADE 사용자는 품사 체계와 품사명 및 인접 조건을 나타내기 위한 자질명들을 임의로 정의하고 형태소 사전을 만들 수 있다. 일단 초기 버전의 형태소 사전이 만들어지면 인접 조건 검사에 따라 올바르게 형태소 분석이 수행되는지 검사하며 이 과정에서 형태소 분석 오류가 발견되면 MADE 프로그램 자체를 수정하는 것이 아니라 형태소 사전에 부가된 인접 조건 정보를 수정·보완하는 식으로 형태소 분석기를 개발해 나갈 수 있다. 이렇게 개발된 형태소 분석기를 다른 응용에서 사용하는 과정에서 분석 오류가 발견되더라도 마찬가지 방법으로 오류를 수정할 수 있다. 이 점은 기존 한국어 형태소 분석기에서 형태소 분석과 관련된 많은 지식들이 프로그램 내부에 인코딩되어 있어 분석 오류가 있거나 분석 결과를 변경하고 싶어도 사용자가 임의로 수정할 수 없는 구조라는 것과 대비가 된다. 한국어 형태소 분석기는 사용자가 임의로 수정할 수 없는 패키지 형태로 보급되던 것과 달리 일본어의 경우에는 사용자가 품사를 정의하고 단어 사이의 연결 관계를 쉽게 수정하여 사용할 수 있는 JUMAN이라는 형태소 분석기가 개발된 바 있다[1, 7].

2. 인접 조건 검사에 의한 형태소 분석

2.1 음절을 경계로 한 분석

기존의 형태소 분석 방법에서는 '**그린다**'와 같은 어절을 분석하기 위해 이것을 '**그리**'와 '**ㄴ다**'로 분리한 후 각각에 대하여 형태소 사전 탐색을 하였다. 또 '**그어라**'와 같은 어절을 분석하기 위해 이것을 '**그**'와 '**어라**'로 분리한 후 여러 단계의 변형 절차를 거쳐 원형인 '**긋**'을 복원해 낸 후 형태소 사전을 탐색하였다. 이러한 작업은 자소 단위의 처리를 요구하며 이를 위해 기존 방법에서는 형태소 분석에 앞서 코드 변환 과정을 반드시 거쳐야 하는 것으로 인식되었다[8]. 이런 절차

는 형태소 분석 속도 개선에 걸림돌이 되는데, 기존 연구에서는 이런 걸림돌은 그대로 둔 채 여러 가지 다른 방법으로 분석 속도를 개선하기 위한 노력을 하였다[9, 10, 11].

하지만 인접 조건 검사에 의한 형태소 분석 방법에서는 코드 변환이나 복잡한 원형 복원 등의 절차를 거치지 않고 음절을 경계로 형태소 분석을 한다. 음절을 경계로 형태소 분석을 한다는 것은 음절을 경계로 사전 탐색을 한다는 것을 의미한다. 예를 들어 '**그린다**'의 경우 기존 방법에서와 같이 '**그리**'와 '**ㄴ다**'로 분리하여 사전 탐색을 하는 것이 아니라 '**그린**'과 '**다**'로 분리하여 즉 음절을 경계로 분리하여 사전 탐색을 하고 형태소 분석을 시도한다는 것이다. '**그어라**'의 경우에도 '**그**'와 '**어라**'로 분리한 후 원형 복원 단계를 거치지 않고 이 상태에서 사전 탐색을 하여 형태소 분석을 한다. 그러면 이제부터 음절을 경계로 한 형태소 분석 방법에 대하여 구체적으로 살펴보기로 한다.

독립언과 같이 하나의 형태소로 구성된 어절은 한 번의 사전 검색으로 형태소 분석을 할 수 있으므로 문제가 되지 않는다. 따라서 여기서는 두 개 이상의 형태소가 결합하여 하나의 어절을 이루는 경우에 대해서만 살펴보자. 먼저 '**집으로**'와 같이 체언과 조사가 결합하여 어절을 이루는 경우를 보자. 사전 탐색을 통해 '**으로**'가 조사이며 '**집**'이 명사라는 것이 밝혀진다면 주어진 어절은 다음과 같이 분석될 것이다[1]. 여기서 NN은 명사, JO는 조사를 나타내는 품사명이다[2].

(1) 집으로 → 집/NN + 으로/JO

이번에는 '**아름답다고**'와 같이 용언에 어미가 결합된 형태의 어절을 예로 들어 보자. 이 경우에도 두 번의 사전 탐색만으로 형태소 분석이 완료되며 그 결과는 아래와 같다.

(2) 아름답다고 → 아름답/AJ + 다고/EM

위에서 살펴본 두 경우 모두 음절을 경계로 한 형태소 분석이 자연스럽게 이루어진다. 하지만 한국어에서는 이처럼 자연스러운 방법으로 형태소 분석을 할 수 없는 경우도 많다. 예를 들어 용언 뒤에 **ㄴ, ㄹ, ㅁ, ㅂ, ㅆ** 등으로 시작하는 어미가 오는 경우나 **아** 또는 **어**로 시작하는 어미가 오는 경우에는 어미의 시작 부분과 용언 어간의 끝부분이 하나의 음절로 결합해 버리기 때문에 용언 어간만 표제어로 등재되어 있는 형태소 사전을 이용할 경우 음절을 경계로 한 형태소 분석이 불가능하게 된다. 다음 절에서는 이러한 경우에 어떻게 음절을 경계로 형태소 분석을 할 수 있는지에 대해서 살펴보기로 한다.

2.2 음운 · 형태 제약 조건

조사 중에는 '**은**'이나 '**는**'과 같이 체언의 마지막 음절이 종성 자음을 가지는지의 여부를 제약하는 것들이 있다. 마찬가지로 어미 중에도 '**아라**'나 '**어라**'와 같이 용언의 마지막 음절이 양성 모음인지 음성 모음인지의 여부를 제약하는 것들이 있다. 이와 같이 왼쪽에 올 수 있는 단어의 음운론적 제약을 가하는 것을 **음운 제약 조건**이라고 한다. 이러한 제약 조건 검사를 위해 형태소 사전을 다음과 같이 구성하였다.

(3) 학교 : ([학교/NN] +모음)
마당 : ([마당/NN] +자음)
보 : ([보/VV] +양성 +모음)
먹 : ([먹/VV] +음성 +자음)

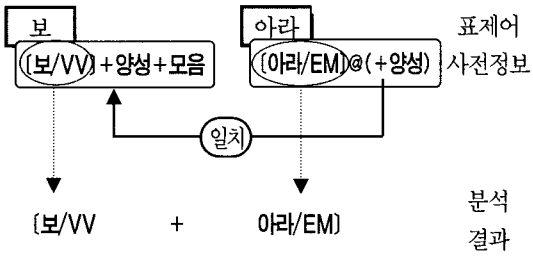
1) 본 논문에서는 어절의 오른쪽 끝에서 시작하여 왼쪽 방향으로 사전 탐색을 하면서 형태소 분석을 하는 것으로 가정하고 있지만 똑같은 방법론을 가지고 반대 방향으로 형태소 분석을 하는 것도 가능하다.

2) 품사 분류 체계 및 명칭은 MADE 사용자가 임의로 정할 수 있지만, 본 논문의 예에서는 중분류 수준의 품사 체계를 가정하고 있으며, NN(명사), NP(대명사), NU(수사), NX(의존명사), DT(관형사), AD(부사), IJ(감탄사), SY(부호), VV(동사), VX(보조동사), SV(동사화접미사), AJ(형용사), AX(보조형용사), SJ(형용사화접미사), CP(계사), JO(조사), EP(선어말어미), EM(어말어미), SN(접미사), PF(접두사) 등과 같은 기호를 사용한다.

(4) 는 : ({는/JO} @(+모음))
은 : ({은/JO} @(+자음))
아라 : ({아라/EM} @(+양성))
어라 : ({어라/EM} @(+음성))

(3)에서 **+모음, +자음, +양성, +음성** 등은 각 표제어의 마지막 음절이 자음으로 끝나는지 모음으로 끝나는지 양성 모음을 가지는지 음성 모음을 가지는지를 나타내는 자질이다. 이러한 자질들이 (4)에서와 같이 **@(…)** 안에서 쓰일 때에는 표제어의 왼쪽에 올 수 있는 단어의 음운 제약 조건을 나타낸다. […]는 표제어에 대한 형태소 분석 결과를 나타내는 부분인데, 이것은 형태소 분석 시 분석 결과를 생성하는 데 사용된다. 형태소 사전이 위와 같이 구성된 경우 인접 조건 검사에 의해 '**보아라**'가 분석되는 과정은 다음과 같다.

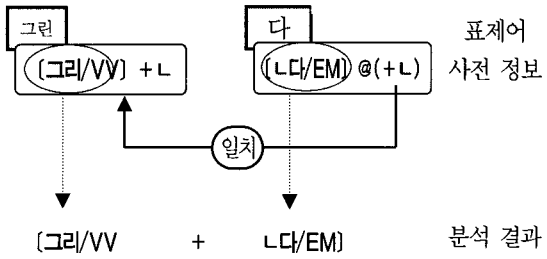
(5)



형태소 사전에서 주어진 어절을 역방향으로 탐색하면 먼저 '**아라**'가 발견된다. 형태소 사전에 **+양성**이란 음운 제약 조건이 명시되어 있으므로 이 단어의 왼쪽에는 양성 모음을 가진 단어만 올 수 있다. 계속해서 음절을 경계로 사전 탐색을 하면 '**보**'가 발견된다. 이 단어에 대한 사전 정보를 보면 **+양성**이라는 음운 정보가 주어져 있으며 이는 '**아라**'에 의해 제기된 음운 제약 조건이 충족된다. 더 이상 분석할 문자열이 없으므로 형태소 사전의 […] 부분을 가지고 와서 결합하면 주어진 어절에 대하여 **{보/VV + 아라/EM}**란 분석 결과를 얻게 된다.

'**그린다**'에서 보듯이 '**ㄴ다**'와 같이 자소를 포함하고 있는 어미는 선행하는 용언의 마지막 음절과 결합해 버린다. 이런 경우 기존 연구에서는 복잡한 원형 복원 과정을 거쳐 용언의 어간 '**그리**'와 어미 '**ㄴ다**'를 찾은 다음 형태소 사전 탐색을 하였다. 그러나 용언의 어간과 어미를 표제어로 등재하는 전통적인 형태소 사전의 개념에서 벗어나 용언 어간에 어미의 자소가 결합된 '**그린**'과 어미에서 자소 부분이 탈락한 '**다**'를 형태소 사전에 표제어로 등재하는 것으로 확대한다면 복잡한 원형 복원 과정을 거치지 않아도 (6)에서 보는 바와 같이 음절을 경계로 한 형태소 분석이 가능해 진다. 이러한 개념의 사전을 **확대된 형태소 사전**이라 한다.

(6)



(6)에서 표제어 '**그린**'에 주어진 **+ㄴ**이라는 자질은 이 표제어가 어간에 **ㄴ**이 결합된 형태임을 나타내는 것이며, 표제어 '**다**'에 주어진 **+ㄴ**은 왼쪽에 **ㄴ**과 결합된 형태가 나와야 함을 나타내는 제약 조건이다. 이와 같이 왼쪽에 올 수 있는 단어의 형태에 제약을 가하는 조건을 **형태 제약 조건**이라고 한다.

'**ㄴ다**' 외에도 '**ㄹ지**', '**ㅂ니다**' 등과 같이 용언 어간의 마지막 음절과 결합할 수 있는 어미들은 모두 같은 방식으로 처리한다[3]. 이를 위해서는 '**그리다**'의 경우 다음과 같은 형태가 확대된 형태

3) **ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아, 어** 등으로 시작하는 어미가 이에 해당한다.

소 사전에 표제어로 등재되어 있어야 한다. 여기서 **+원형**은 표제어가 용언 어간임을 나타내는 자질이며, **+ㄴ**, **+ㄹ**, **+ㅁ**, **+ㅂ**, **+ㅆ**, **-아**는 표제어가 용언 어간에 **ㄴ**, **ㄹ**, **ㅁ**, **ㅂ**, **ㅆ**, **아**가 결합된 형태임을 나타내는 자질이다.

(7) 그리 : ([그리/VV] +원형 +양성 +모음)
그린 : ([그리/VV] +ㄴ)
그릴 : ([그리/VV] +ㄹ)
그림 : ([그리/VV] +ㅁ)
그립 : ([그리/VV] +ㅂ)
그렸 : ([그리/VV + 었/EP] +ㅆ)
그려 : ([그리/VV] -아 +양성)

용언 중에는 뒤에 오는 어미에 따라 어간 일부가 변하는 불규칙 용언이 있다. 예를 들어, '**듣다**', '**긋다**', '**돕다**' 등과 같은 불규칙 용언은 특정 어미 앞에서 '**들어서**', '**그어서**', '**도와서**'와 같이 어간의 일부가 변하게 된다. 전통적인 형태소 분석 방법에서는 이런 경우 복잡한 원형 복원 절차를 거쳐야 하는 것으로 되어 있다. 그러나 인접 조건 검사에 의한 형태소 분석 방법에서는 불규칙 용언의 경우에도 별도의 원형 복원 절차를 거치지 않고 음절을 경계로 한 형태소 분석을 할 수 있다. 단, 인접 조건 검사만으로 불규칙 용언을 분석하려면 (7)에서와 같이 용언 어간뿐만 아니라 어미의 일부가 결합된 형태도 확대된 형태소 사전에 표제어로 등재되어 있어야 한다.

다음은 불규칙 용언을 확대된 형태소 사전에 등재한 예이다. '**들**'이나 '**그**'는 '**들어**', '**그어**' 등과 같이 '**아**'나 '**어**'로 시작하는 어미 앞에서 쓰일 수 있으므로 **+아**라는 자질을 가진다. 그런데 '**도와도**', '**미워도**'에서 볼 수 있듯이 **ㅂ** 불규칙 활용을 하는 용언은 '**아도**'나 '**어도**'와 결합하지 않고 '**와도**'나 '**워도**'와 결합한다. 이와 같이 '아'나 '어' 대신 '**와**'나 '**워**'로 시작하는 어미 앞에서만 쓰일 수 있는 형태에 대해서는 **+아** 대신에 **+와**라는 자질을 주어 구분하였다.

(8) 듣 : ([듣/VV] +원형 +음성)
들 : ([듣/VV] +아 +음성)
([들/VV] +아 +음성)
돕 : ([돕/VV] +원형 +양성)
도 : ([돕/VV] +와 +양성)
긋 : ([긋/VV] +원형 +음성)
그 : ([긋/VV] +아 +음성)

(9)는 확대된 형태소 사전의 어미 부분을 예시한 것이다. '**와도**'나 '**워도**'는 어미가 아니므로 전통적인 형태소 사전에는 표제어로 등재되지 않았지만 인접 조건 검사에 의한 형태소 분석을 위한 확대된 형태소 사전에서는 표제어로 등재된다. 다만 '**와도**'나 '**워도**'는 **ㅂ** 불규칙 활용을 하는 용언과 결합될 수 있다는 것을 나타내기 위해 **+아**라는 자질 대신 **+와**라는 자질이 사용되었다.

(9) 고 : ([고/EM] @(+원형))
도 : ([아도/EM] @(-아 +양성))
([어도/EM] @(-아 +음성))
아도 : ([아도/EM] @(+아 +양성))
어도 : ([어도/EM] @(+아 +음성))
와도 : ([아도/EM] @(+와 +양성))
워도 : ([어도/EM] @(+와 +음성))

(8)-(9)와 같은 확대된 형태소 사전을 이용하면 '**듣고**', '**돕고**', '**긋고**', '**들어도**', '**도와도**', '**그어도**' 등과 같은 불규칙 활용을 하는 용언이 포함된 어절을 올바르게 분석할 수 있다. 여기서는 **ㄷ**, **ㅂ**, **ㅅ** 불규칙 용언만 예로 들었는데 다른 유형의 불규칙 활용을 하는 용언도 마찬가지 방법으로 분석할 수 있다. '**으**' 매개 모음 문제도 유사한 방법으로 해결할 수 있다.

2.3 품사 제약 조건

조사는 체언과 함께 사용되지만 보조사 '**는**', '**도**', '**만**' 등은 체언뿐만 아니라 부사와 함께 사

용될 수도 있다. 또한 어미 중에는 모든 용언에 두루 쓰이는 것이 있는가 하면 동사나 형용사에 한해서 사용되는 것도 있다. 예를 들어 어미 '**거나**'는 동사나 형용사 어느 것과도 결합할 수 있는 반면, 어미 '**는군요**'는 동사와만 결합할 수 있으며 '**군요**'는 형용사와만 결합할 수 있다. 이와 같이 조사나 어미는 자신의 왼쪽에 오는 형태소의 품사에 제약을 가하는 것을 볼 수 있는데 이러한 현상은 조사나 어미뿐만 아니라 다른 품사에서도 발견된다. 예를 들어 체언의 경우 자신의 왼쪽에 임의의 품사가 올 수 있는 것이 아니라 체언, 접두사, 관형사 또는 부호 등만 올 수 있는 것으로 제약한다. 따라서 이것을 "임의의 형태소는 자신의 왼쪽에 오는 형태소의 품사를 제약할 수 있다"라고 일반화시킬 수 있는데, 이러한 제약 조건을 **품사 제약 조건**이라고 한다. (10)은 품사 제약 조건이 명시된 형태소 사전의 예이다. 여기서 #(…) 부분은 품사 제약 조건을 나타낸다[4].

(10) 에서 : ([에서/JO] #(NN NP NU NX NF))
은 : ([은/JO] #(NN NP NU NX NF AD EM JO) @(+자음))
는 : ([는/JO] #(NN NP NU NX NF AD EM JO) @(+모음))
거나 : ([거나/EM] #(VV VX SV AJ AX SJ EP) @(+원형 | +ㅆ))
는군요 : ([는군요/EM] #(VV VX SV EP) @(+원형))
군요 : ([군요/EM] #(AJ AX SJ EP) @(+원형 | +ㅆ))

위의 예에서 보듯이 조사 '**에서**'는 명사, 대명사, 수사, 의존 명사와 같은 체언과 결합하지만, 조사 '**은**'이나 '**는**'은 체언은 물론 부사와도 결합하는 것으로 되어 있다. 그런데, '**은**'이나 '**는**'은 음운·형태 제약 조건도 가지고 있으므로 아무 체언이나 부사와 결합할 수 있는 것이 아니라 지정된 음운·형태 제약 조건을 만족하는 체언이나 부사하고만 결합할 수 있다. 또, 어미 '**거나**'는 용언이나 선어말 어미와 결합하는 반면 '**는군요**'는 동사나 선어말 어미와, '**군요**'는 형용사나 선어말 어미와 결합하는 것으로 되어 있다. 그런데 '**거나**'나 '**군요**'는 **+원형** 또는 **+ㅆ** 이라는 형태 제약 조건을 가지고 있으므로 이 단어의 왼쪽에는 용언 어간인 '**사**', '**돕**' 등이나 **+ㅆ** 형태 정보를 가진 '**샀**', '**았**', '**었**' 등이 올 수 있다.

3. 형태소 사전

3.1 원시 형태소 사전과 확대된 형태소 사전

2절에서 소개한 인접 조건 검사에 의한 형태소 분석 방법은 매우 단순하여 효과적으로 형태소 분석을 할 수 있다. 하지만 여기에서 사용되는 확대된 형태소 사전은 순수한 형태소뿐만 아니라 어미의 일부가 결합된 복합 형태까지도 표제어로 가지고 있어야 하기 때문에 다른 방법론에서 사용되던 단순한 형태의 형태소 사전에 비하여 구축하기가 어렵다는 문제가 있다. 이 문제를 해결하기 위하여 MADE에서는 단순한 구조의 원시 형태소 사전으로부터 인접 조건 검사에 의한 형태소 분석에 필요한 정보를 제공하는 복잡한 구조의 확대된 형태소 사전을 자동 생성하는 기능을 제공한다.

원시 형태소 사전은 구조가 매우 단순하여 만들기가 어렵지 않다. 그러면 이제부터 원시 형태소 사전의 구조를 살펴보기로 한다. 독립언이나 체언류 등은 원시 형태소 사전에 간략하게 품사 정보만 명시하면 된다. 그러면 MADE에 내장된 사전 변환 모듈에 의해 인접 조건에 의한 형태소 분석을 하는데 필요한 음운 정보와 품사 제약 조건 등이 자동으로 부가되어 확대된 형태소 사전이 만들어 진다. 음운 정보는 한국어의 음절을 분

4) 이는 설명 목적상 음운·형태 제약 조건과 품사 제약 조건을 구분하기 위한 것으로 MADE에서는 두 제약 조건 공히 @(…) 형식으로 나타낸다.

석하여 기계적으로 부가할 수 있으며, 품사 제약 조건은 MADE 사용자가 미리 정의해 둔 디폴트 품사 제약 조건에 따라 주어진다[5].

용언류도 체언류와 비슷하게 원시 형태소 사전을 만들면 된다. 용언 중에는 불규칙 활용을 하는 것들도 있는데, 이런 경우에는 (11)에서 보듯이 불규칙 활용 정보도 명시해 주어야 한다[6].

(11) (가) 아름답 : (AJ +ㅂ불)
(나) 긋 : (VV +ㅅ불)
(다) 푸 : (VV +우불)

용언류의 경우에는 형태소 분석을 하기 위해 2절에서 본 것처럼 ㄴ, ㄹ, ㅁ, ㅂ, ㅆ, 아, 어 등이 결합된 형태가 필요하므로 (11)과 같은 원시 형태소 사전은 사전 변환 모듈에 의해 (12)와 같은 형태로 확대된다[7].

(12) (가) 아름답 : ([아름답/AJ] +원형 +양성 +자음)
아름다 : ([아름답/AJ] +양성 +와)
(나) 긋 : ([긋/VV] +원형 +음성 +자음)
그 : ([긋/VV] +음성 +아 +으)
(다) 푸 : ([푸/VV] +원형 +음성 +모음)
푼 : ([푸/VV] +ㄴ)
풀 : ([푸/VV] +ㄹ)
품 : ([푸/VV] +ㅁ)
풉 : ([푸/VV] +ㅂ)
퍼 : ([푸/VV] +음성 -아)
펐 : ([푸/VV] +음성 +ㅆ)

조사나 어미의 경우에는 인접 조건이 다소 복잡한데, 원칙적으로 원시 형태소 사전에 이러한 인접 조건을 일일이 다 명시해 주어야 한다. 예를 들어 조사의 경우 체언류 뒤에 쓰일 수 있는 것도 있고 보조사와 같이 체언류 외에 부사나 다른 조사나 어미 뒤에서도 쓰일 수 있는 것도 있다. 이런 제약 조건들은 원시 형태소 사전에서 (13)과 같이 기술될 것이다.

(13) 에서 : ([에서/JO] #(NN NP NX NU NF) +모음)
은 : ([은/JO] #(NN NP NX NU NF AD EM JO) @(+자음) +자음))
는 : ([는/JO] #(NN NP NX NU NF AD EM JO) @(+모음) +자음))

그런데 모든 조사에 대하여 이런 제약 조건들을 일일이 기술하기란 여간 번거로울 뿐만 아니라 실수로 한두 개를 빠뜨리면 형태소 분석이 올바르게 이루어지지 않게 된다. 이런 실수를 미연에 방지하고 사용자가 능률적으로 형태소 사전을 구축할 수 있도록 하기 위하여 MADE에서는 매크로란 개념을 사용한다. 매크로를 잘 정의해 두면 원시 형태소 사전을 구축할 때 (13)과 같이 복잡하게 기술하지 않고 (14)와 같이 간결하게 기술할 수 있다[8].

(14) 에서 : ([에서/JO] @체언 +모음)
은 : ([은/JO] @자음|보조사 +자음)
는 : ([는/JO] @모음|보조사 +자음)

5) '척', '대', '개' 등과 같이 물건의 개수를 헤아릴 때 사용되는 명사의 왼쪽에 올 수 있는 단어의 품사는 보통 명사의 왼쪽에 올 수 있는 단어의 품사와 다소 차이가 있다. 이런 특별한 경우를 제외하면 통일 품사의 단어들은 자신의 왼쪽에 올 수 있는 단어의 품사가 대체로 일치한다. 따라서 원시 형태소 사전을 만들 때 개별 단어의 품사 제약 조건을 일일이 명시할 것이 아니라 일반적으로 적용될 수 있는 품사 제약 조건을 정의하고 따로 명시된 품사 제약 조건이 없는 경우 이것이 적용되도록 하는 것이 효과적이다. 이런 목적으로 정의된 품사 제약 조건을 디폴트 품사 제약 조건이라고 한다.

6) 용언의 불규칙 활용에 대해서는 [12]에 자세히 소개되어 있다.

7) 여기에서 사용된 자질명은 예시이며 사용자에 따라 다른 자질명을 정의하여 사용할 수 있다.

8) 이 예에서 @체언은 #(NN NP NX NU NF)로 정의된 매크로이며, @자음|보조사와 @모음|보조사는 각각 #(NN NP NX NU NF AD EM JO) @(+자음)와 #(NN NP NX NU NF AD EM JO) @(+모음)로 정의된 매크로이다.

추후 MADE에 의해 원시 형태소 사전이 확대된 형태소 사전으로 변환될 때 매크로들은 미리 정의된 내용으로 자동 확장될 것이다.

어미의 경우도 조사와 마찬가지로 매크로를 사용하면 형태소 원시 형태소 사전을 간결하게 만들 수 있다. 다음은 매크로를 사용하여 어미 사전을 만든 예이다.

(15) 거나 : ([ㄹ거나/EM] @ㄹ|용언 +원형)
니다 : ([ㅂ니다/EM] @ㅂ|용언 +원형)
으셔야 : ([시/EP + 어야/EM]
@자음|용언 +원형 +연결)
셔야 : ([시/EP + 어야/EM]
@모음|용언 +원형 +연결)

여기서 **@ㄹ|용언**, **@ㅂ|용언**, **@자음|용언**, **@모음|용언**은 각각 표제어가 **ㄹ** 또는 **ㅂ**과 결합된 형태와 함께 쓰일 수 있는 (부분) 어미, 자음 또는 모음으로 끝나는 용언과 쓰일 수 있는 어미임을 나타내는 매크로들이다. 물론 이러한 매크로들은 나중에 적절한 인접 조건으로 확장될 수 있도록 사전에 정의되어 있어야 한다.

3.2 보다 발전된 형태소 사전 구축

지금까지는 MADE 사용자가 구축해야 할 원시 형태소 사전의 일반적인 내용에 대하여 살펴보았다. 이 절에서는 인접 조건 검사에 의한 형태소 분석 방법을 활용하여 보다 발전된 형태소 분석기를 만드는 몇 가지 사례를 살펴보기로 한다.

실제 문서를 보면 띄어쓰기가 잘못 된 사례를 많이 볼 수 있다. 띄어쓰기가 잘못 된 어절에 대한 형태소 분석은 쉽지 않지만, 형태소 사전 구축 시 약간의 노력만 하면 전형적인 띄어쓰기 오류 어절도 분석이 가능하도록 할 수 있다. 예를 들어 **'먹은것'**, **'먹는것'**, **'먹을것'**, **'먹는수'**, **'먹을수'**와 같이 관형사형 어미가 있는 어절 뒤에 의존 명사 **'수'**가 오는 경우 띄어쓰기를 무시하는 경우가 많은데, 이러한 어절도 (16)과 같이 형태소 사전을 작성하면 분석이 가능하게 된다. 여기서 **+는**은 **'는'**과 같은 관형형 어미에 부여되는 자질이다.

(16) 것 : ([것/NX] #(EM)
@(+는 | +ㄴ | +ㄹ))
수 : ([수/NX] #(EM) @(+는 | +ㄹ))

다른 예로 **'천사와같은'**, **'별과같은'**, **'아름답게 되어'**에서와 같이 띄어쓰기가 잘못 된 어절도 다음과 같이 형태소 사전만 구축해 준다면 아무런 문제없이 형태소 분석이 가능하게 된다.

(17) 와같 : ([와/JO + 같/AJ] #(NN NP
NU NX) @(+모음) +원형)
과같 : ([과/JO + 같/AJ] #(NN NP
NU NX) @(+자음) +원형)
게되 : ([게/EM + 되/VX] #(VV SV
AJ SJ EP) @(+원형) +원형)

명사에 **'하다'**와 같은 동사화 접미사가 붙어 명사에서 동사로 파생되는 경우가 있다. 동사화 접미사 **'하다'** 앞에 명사가 올 수 있음을 나타내기 위하여 (18)과 같이 품사 제약 조건을 줄 수 있다.

(18) 하 : ([하/SV] #(NN))

형태소 사전을 이렇게 만들면 **'부하는'**이 **[부/NN + 하/SV + 는/EM]**으로 분석될 수도 있다. 이와 같이 분석되는 것을 막으려면 **'하다'** 앞에 아무런 명사나 올 수 있는 것이 아니라 가령 동작성 명사만 나올 수 있는 것으로 제약을 가하면 될 것이다. 이런 제약을 가하기 위해서 **+동작**이란 자질을 정의하고, 동작성 명사에 이 자질을 부여하고 (18)을 (19)와 같이 수정하면 될 것이다.

(19) 하 : ([하/SV] #(NN) @(+동작))

이 예에서 보듯이 자질은 음운·형태 조건을 기술하는 데에만 사용되는 것이 아니라 사용자의 필요에 따라 여러 가지 용도로 사용될 수 있다.

기존의 형태소 분석기에서는 분석 복잡도를 줄이기 위하여 복합 조사나 어미를 형태소 사전에 하나의 표제어로 등재하는 경우가 많았다[4]. 하지만 이에 대하여 분석기의 성능보다는 국어학적 내용을 중시하는 측에서는 기존 형태소 분석기가 복합 조사나 어미의 원형을 제대로 복원하지 않음을 지적하였다[13]. 2 절에서 설명한 것처럼 MADE의 형태소 사전에서 〔…〕 부분은 형태소 분석 결과를 생성할 때 사용된다. 따라서 복합 조사나 어미에 대하여 (20)과 같이 형태소 사전을 구축한다면 복합 조사나 어미를 하나의 표제어로 등재하여 분석 복잡도를 줄이면서도 복합 조사나 어미의 원형을 제대로 복원해 내지 못 한다는 비난을 피할 수 있다.

(20) 에서부터는 : ((〔에서/JO + 부터/JO + 는/JO〕 @체언 +자음)
라고까지는 : ((〔라고/EM + 까지/JO + 는/JO〕 @용언 +자음)

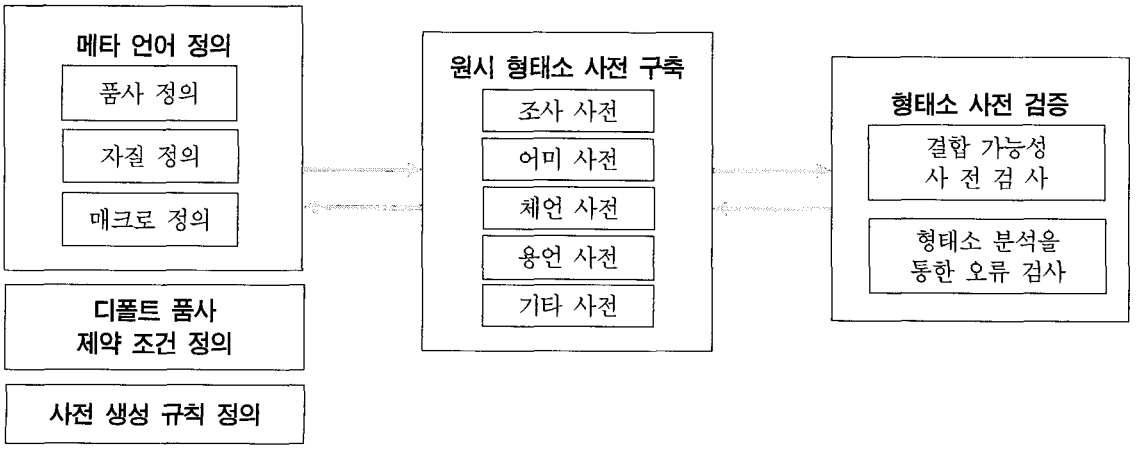
물론 형태소 분석기의 용도에 따라 복합 조사나 어미를 단위 조사나 어미로 분리하지 않는 것이 더 좋을 수도 있다. 이런 경우에는 (21)과 같이 형태소 사전을 구축하면 된다.

(21) 에서부터는 : ((〔에서부터는/JO〕 @체언 +자음)
라고까지는 : ((〔라고까지는/EM〕 @용언 +자음)

4. 형태소 분석기 개발 및 배포

MADE 사용자는 (그림 1)과 같은 절차에 따라 형태소 분석기를 개발한다. 우선 메타 언어를 정의하고 이어 디폴트 품사 제약 조건과 사전 생성 규칙을 정의한다. 메타 언어 정의 단계에서는 품사 체계를 수립하고 적절한 품사 이름을 정의한다. 자질은 기본적으로 본 논문에서 사용된 것들을 그대로 사용하여도 되지만 필요에 따라 다른 이름이나 새로운 자질을 추가로 정의하여 사용할 수 있다. 매크로는 형태소 사전을 만들기 전에 미리 정의할 수도 있지만 대개는 사전 구축 과정에서 매크로의 필요성이 인식되기 때문에 사전을 만들어 나가는 동안 점진적으로 추가하는 경우가 보통이다.

디폴트 품사 제약 조건은 특정 품사의 단어 왼쪽에 일반적으로 올 수 있는 품사들의 목록이다.



(그림 1) MADE를 이용한 형태소 분석기 개발 과정

예를 들어 명사 왼쪽에는 일반적으로 다른 명사류나 관형사, 각종 부호, 접두사 등이 나올 수 있으므로 명사 NN에 대한 디폴프 품사 제약 조건은 (NN NU NX DT PF)이 된다. 이러한 품사 제약 조건은 각 품사 별로 정의해야 한다.

사전 생성 규칙은 용언에 대해서만 적용되는 규칙으로 MADE에 내장된 사전 변환 모듈에 의해 (11)과 같은 원시 형태소 사전으로부터 (12)와 같은 확대된 형태소 사전이 생성될 때 각 형태에 대해 어떤 자질들을 부여할지 결정할 때 사용된다. 사전 생성 규칙은 용언의 규칙 또는 불규칙 유형별로 정의하여야 한다.

메타 언어와 디폴트 품사 제약 조건, 사전 생성 규칙은 쉽게 정의할 수 있지만 형태소 사전 구축은 시간이 좀 걸린다. 하지만 인터넷 등에 공개된 기본적인 어휘 목록을 이용한다면 형태소 사전 구축 시간을 상당히 단축시킬 수는 있다. 3절에서 설명한 바와 같이 원시 형태소 사전을 만들 때 체언 사전과 기타 사전의 경우에는 품사 정보만, 용언 사전의 경우에는 품사 정보와 불규칙 활용 정보만 명시하면 되기 때문에 이들 품사에 대한 원시 형태소 사전을 만드는 것은 그다지 어렵지 않다. 조사와 어미는 쓰임새가 비교적 복잡하므로 조사와 어미 사전에는 그에 상응하는 인접 조건을 일일이 명시해 주어야 하는 어려움이 있다. 하지만 조사나 어미는 어휘 수도 많지 않고 **'으려'**, **'으려고'**, **'으려면'**, **'으려던'**과 같이 형태론적 특성이 유사한 조사나 어미는 동일한 인접 조건을 가지므로 실제 작업량은 그다지 많지 않다.

초기 버전의 형태소 사전을 구축하고 나면 MADE를 이용하여 바로 형태소 분석을 수행해 볼 수 있다. 따라서 이후부터는 실제로 형태소 분석을 수행하면서 형태소 사전의 오류를 수정·보완하는 작업을 한다. 직접 문장을 입력하여 형태소 사전의 오류를 찾아 나갈 수도 있으나 MADE에서 제공하는 결합 가능성 사전 검사 기능을 이용한다면 보다 효과적으로 형태소 사전의 오류를 수정할 수 있다. 이 과정을 거치면 그 다음은 실제 텍스트를 대상으로 형태소 분석을 수행하여 오류를 찾아 형태소 사전을 수정·보완하는 작업을 반복한다. 형태소 사전을 한 번에 정확하게 만들기도 어렵고 또 실제 문장을 분석하다 보면 미처 생각하지 못한 여러 가지 예외적인 경우도 발생하기 때문에 형태소 사전의 오류를 수정·보완하는 작업을 얼마나 많이 했는가는 형태소 분석기의 정확도와 직접적으로 관련이 있다.

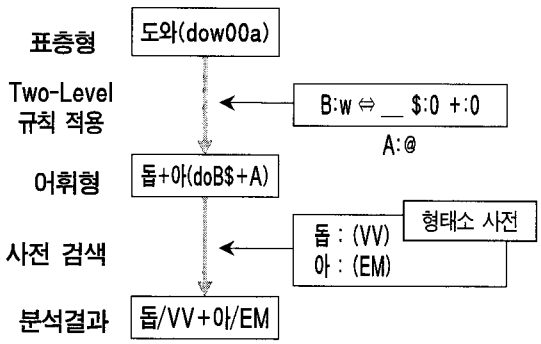
형태소 분석 결과를 유형별로 구분하여 볼 수 있는 기능도 가지고 있다. 이 기능은 형태소 사전을 구축한 후 실제로 형태소 분석을 하면서 형태소 사전의 오류를 고쳐 나가는 과정에서 자주 사용된다. 이 기능을 이용하면 분석에 성공한 어절만 골라서 보거나 반대로 실패한 어절만 골라서 볼 수도 있고, 어절 길이를 지정한다든지 형태소 분석 개수를 지정한다든지 해서 조건에 맞는 어절의 형태소 분석 결과만 골라 볼 수도 있다. 즉 분석에 실패한 어절만 골라서 관찰함으로써 공통된 실패 원인이 있는지 살펴 볼 수도 있고, 길이를 지정하여 길이가 긴 어절만 따로 골라서 분석 결과의 타당성을 따져 볼 수도 있다. 품사 패턴을 지정하고 이 패턴이 포함 된 분석 결과만 골라서 보는 기능도 있다.

MADE는 내부에 언어 독립적인 형태소 분석 엔진을 가지고 있다. 이 엔진은 MADE.DLL이라는 라이브러리 형태로 존재하며 MADE와 분리하여 독립적으로 사용할 수도 있다. 형태소 분석기 개발 단계에서는 GUI (Graphical User Interface) 형식의 MADE를 사용하여 작업을 하게 되지만 일단 개발이 끝나 독립된 형태의 형태소 분석기로 배포할 때에는 MADE.DLL과 확대된 형태소 사전 및 메타 언어 정의 부분만 배포하면 된다. MADE.DLL은 확대된 형태소 사전에서 인접 조건을 가지고 오고 이를 기반으로 인접 조건 검사를 수행함으로써 형태소 분석을 하는 모듈로 내부에 한국어 형태소 분석과 관련된 언어적인 사항은 전혀 인코딩되어 있지 않다. 따라서 사용 과

정에서 분석 오류가 발견되더라도 MADE.DLL을 수정하는 것이 아니라 형태소 사전만 수정하면 오류를 없앨 수 있는 구조이다.

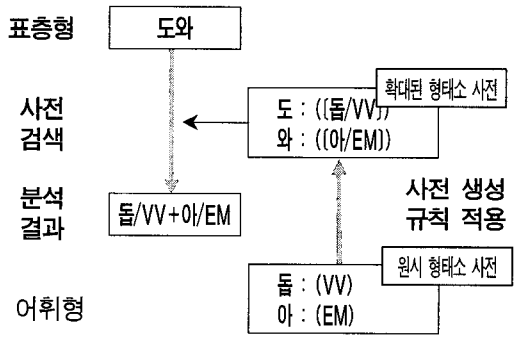
5. PC-KIMMO와의 방법론 비교

PC-KIMMO에서 사용하는 Two-Level 형태소 분석 모델에서는 (그림 2)와 같이 Two-Level 규칙을 적용하여 표층형을 어휘형으로 직접 변환하고 이렇게 해서 얻어진 어휘형으로 형태소 사전을 탐색하는 방법으로 형태소 분석을 한다[9]. 여기서 사용되는 형태소 사전에는 어휘형만 등재되어 있다. 이 모델에서는 입력 어절에 대하여 매번 Two-Level 규칙을 적용하여 표층형에서 어휘형으로 변환하는 절차를 거쳐 형태소 분석을 한다.



(그림 2) Two-Level 형태소 분석 모델

인접 조건 검사에 의한 형태소 분석은 Two-Level 형태소 분석 모델에서와 같이 표층형과 어휘형만 인정한다는 점에서 서로 유사하다 할 수 있다. 하지만 두 가지 중요한 차이점이 있다. 먼저 Two-Level 모델에서는 형태소 분석 중에 표층형을 어휘형으로 변환한 후 어휘형으로 형태소 사전을 검색하는 데 반하여 인접 조건 검사에 의한 형태소 분석에서는 아무런 변환없이 표층형으로 직접 형태소 사전 검색을 한다. 형태소 분석 중에 아무런 변환을 하지 않는 대신 확대된 사전이라는 별도의 사전을 둔다. 이 사전은 사전 생성 규칙을 적용하여 Two-Level 모델과는 역방향인 어휘형에서 표층형으로 변환하여 얻은 것이다. 이처럼 형태소 분석 중에 아무런 변환을 하지 않고 사전 검색만으로 분석을 수행하므로 인접 조건 검사에 의한 형태소 분석은 빠른 형태소 분석을 하는 데 적합한 모델이다. (그림 3)은 인접 조건 검사에 의한 형태소 분석 방법을 나타낸 것이다. 이 그림에서는 편의상 인접 조건 등은 나타내지 않았다.



(그림 3) 인접 조건 검사에 의한 형태소 분석 모델

[5]에서 지적된 바와 같이 Two-Level 규칙은 작성하기도 까다롭고 작성된 규칙을 이해하기도 쉽지 않다. 더구나 Two-Level 형태소 분석 모델을 구현한 PC-KIMMO에서는 직접 한글 입출력이 되지 않으므로 한글을 영어 알파벳으로 바꾸어 표현하기 위한 별도의 프로그램을 작성하거나[5], PC-KIMMO의 입출력 부분을 확장하여 한글을 다룰 수 있도록 해야 하는 불편함이 있다[6]. 이에 반하여 MADE는 한국어 형태소 분석기 개발을 위해 만들어진 도구로서 사용하기 편리할 뿐만 아니라 사전 생성 규칙은 하나의 어휘형에서 변환될 수 있는 표층형에 대하여 적절한 자질 정보만 부가하면 되는 정도로 간단하다. 하나의 어휘형으로부터 변환되어 생길 수 있는 표층형은 MADE에 내장된 사전 변환 모듈에 의해 자동으

9) 실제로 Two-Level 모델에서 사용되는 형태소 사전은 훨씬 복잡하지만 여기서는 단순화하여 나타내었다.

로 생성된다.

PC-KIMMO는 한국어 형태소 분석기 개발에 편의를 제공하는 기능이 없다. 하지만 MADE에서는 용언과 어미의 결합 가능성을 미리 테스트해 볼 수 있는 기능이 있어 실제 텍스트를 대상으로 형태소 분석을 해 보지 않더라도 사전에 등록된 용언과 어미에 대해 올바르게 형태소 분석이 이루어지는지를 검사하고 오류가 있는 경우 형태소 사전을 수정·보완할 수 있는 편리함이 있다. 또 실제 텍스트를 대상으로 형태소 분석을 하는 경우 여러 가지 유형별로 구분하여 분석 결과를 볼 수 있어 보다 편리하게 오류를 찾아 볼 수 있다.

MADE는 비록 한국어 형태소 분석기 개발을 목적으로 만들어진 도구이지만 MADE에 내장된 형태소 분석 엔진은 언어 독립적이므로 원칙적으로 한국어 이외의 다른 언어에 대해서도 적용해 볼 수 있다. 다만 한국어의 경우 사전 변환 모듈이 내장되어 있어 원시 형태소 사전에서 확대된 형태소 사전을 자동 생성할 수 있지만 다른 언어에 적용할 경우에는 확대된 사전을 수작업으로 만들어야 하는 불편함은 감수해야 한다.

6. 결론

MADE는 실용적인 한국어 형태소 분석기를 직접 만들어 볼 수 있는 기능을 제공하는 개발 도구이다. 이 도구에서는 인접 조건 검사만으로 형태소 분석을 수행하는 아주 단순한 알고리즘을 사용하고 있다. 이 알고리즘은 언어 독립적이며 형태소 분석에 필요한 모든 정보가 형태소 사전에서 제공됨을 전제로 한다. 따라서 MADE 사용자는 별도의 프로그래밍을 하지 않더라도 형태소 사전을 만들기만 하면 형태소 분석기를 얻을 수 있다. MADE는 형태소 사전을 구축하고 형태소 분석 오류를 찾아서 수정·보완하는 작업을 보다 편리하게 할 수 있는 기능들을 제공한다. 형태소 분석에 필요한 모든 정보가 형태소 사전에서 제공되므로 추후 분석 오류가 발견되더라도 형태소 분석기 프로그램을 수정할 필요는 전혀 없으며, 형태소 사전에서 오류와 관련된 부분을 수정하기만 하면 되는 구조이므로 유지 및 보수가 용이하다.

MADE에 내장된 형태소 분석 엔진은 독립적으로 사용될 수도 있다. 따라서 MADE를 사용하여 형태소 분석기를 만든 후 MADE 없이 독자적으로 사용할 수 있는 형태소 분석기를 만드는 것도 가능하다. 이 독립적인 형태소 분석기는 형태소 분석기를 필요로 하는 다른 응용 소프트웨어에 내장되어 사용될 수 있다.

참 고 문 헌

[1] Hammarstrom, Harald, "Overview of IT-based tools for learning and training grammar", Project Report, Department of Linguistics, Uppsala University, pp.9-15, 2002.

[2] Koskenniemi, Kimmo, "Two-level model for morphological analysis", Proceedings of the 8th International Joint Conference on Artificial Intelligence, pp.683-685, 1983.

[3] Antworth, Evan, "PC-KIMMO: A Two-Level Processor for Morphological Analysis", Occasional Publications in Academic Computing No. 16, Summer Institute of Linguistics, 1990.

[4] Kwon, Hyuk-Chul, Young-Soog Chae, "A Dictionary-Based Morphological Analysis", Proceedings of Natural Language Processing, Pacific Rim Symposium '91, Singapore, pp.178-185, 1991.

[5] Jang, Tae-Yeoub, "A Two-level Morphological Analysis of Korean", Proceedings of the Postgraduage Conference, Department of Linguistics and Applied Linguistics, The University of Edinburgh, 1998.

[6] Kim, Deok-Bong, Sung-Jin Lee, Key-Sun Choi

and Gil-Chang Kim, "A Two-Level Morphological Analysis of Korean", Proceedings of the 15th Conference on Computational Linguistics, pp.535-594, 1994.

[7] Matsumoto, Yuji and Makoto Nagao, "Improvements of Japanese Morphological Analyzer JUMAN", Proceedings of the International Workshop on Sharable Natural Language Resources, pp.22-28, 1994.

[8] 강승식, 한국어 형태소 분석과 정보 검색, pp.73-247, 홍릉과학출판사, 2002.

[9] 임희석, 윤보현, 임해창, "배제 정보를 이용한 효율적인 한국어 형태소 분석기", 한국정보과학회 논문지, 제22권 제6호, pp.957-964, 1995.

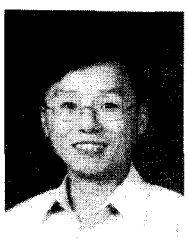
[10] 최재혁, 이상조, "양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 횟수 감소 방안", 정보과학회논문지, 20권, 10호, pp.1497-1507, 1993.

[11] 김영관, 박민식, 최진석, 권혁철, "사전 성능 개선을 통한 한국어 형태소 분석기의 분석 속도 향상", 제11회 한글 및 한국어 정보처리 학술대회 논문집, pp.479-483, 1999.

[12] 남기심, 고영근, 표준 국어 문법론, pp.40-230, 탑출판사, 1997.

[13] 은종진, 박선영, "고성능 한국어 형태소 분석을 위한 어미 분류", 제12회 한글 및 한국어 정보처리 학술대회 논문집, pp.41-47, 2000.

◐ 저 자 소 개 ◑



심 광 섭(Kwangseob Shim)
1986년 서울대학교 컴퓨터공학과 졸업 (학사)
1988년 서울대학교 대학원 컴퓨터공학과 졸업 (석사)
1994년 서울대학교 대학원 컴퓨터공학과 졸업 (박사)
1995년~현재 성신여자대학교 컴퓨터정보학부 교수
관심분야 : 자연어처리, 한국어정보처리, 형태소분석, 구문분석, 인공지능
E-mail : shim@sungshin.ac.kr