한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능

강 승 식
한성대학교 정보전산학부
kang@ham.hansung.ac.kr

이 하 규
한림대학교 전산정보학과
hglee@sun.hallym.ac.kr

Morphological Analysis and Spelling Check Function of Korean Morphological Analyzer HAM

Kang, Seung-Shik
Hansung University

Lee, Hagyu
Hallym University

요 약

한국어 형태소 분석기의 효율성에 영향을 미치는 요인은 분석 알고리즘의 효율성보다도 어휘 사전 등 형태소 분석과 관련된 여러 가지 요인들이 미치는 영향이 훨씬 더 크다. 따라서 단어의 유형 분류 기법이나 불규칙 용언의 분석 방법을 비롯하여 어휘 사전의 구조 및 크기, 알고리즘의 선택과 구현 등 형태소 분석과 관련된 모든 요소들을 형태소 분석에 적합하도록 구성하여야 한다. 본 논문에서는 어휘형태소 사전과 문법형태소 사전의 크기, 한글 문서에 나타나는 단어의 특성 등 형태소 분석기의 효율 및 성능에 영향을 미치는 요소들을 고찰하였다. 그 결과로 알고리즘의 효율보다는 사전 탐색 시간이 형태소 분석에 미치는 영향이 매우 크다는 것을 알 수 있었다. 이와 같이 형태소 분석기의 성능에 영향을 미치는 요인들을 고려하여 구현된 범용 형태소 분석기 HAM에 대하여 형태소 분석 기능과 철자 검사 기능을 실험하였다. 형태소 분석 성공률에 대한 실험 결과 99.46%의 분석률을 보이고 있으며, 맞춤법 검사 기능으로는 상용화된 철자 검사기와 비슷한 성능을 보이고 있다. HAM의 처리 속도는 pentium 120MHz linux 2.0 환경에서 1 초에 약 1,000 단어를 분석한다.

1. 서 론

형태소 분석은 어휘사전과 형태론적 변형 규칙을 이용하여 입력 문자열에 대해 형태론적 변형과 형태소 분리 문제를 처리하는 과정이다. 형태소 분석은 어휘사전을 기반으로 하기 때문에 어휘사전을 구성하고, 사전에 수록되는 형태소에 관한 기준이 정해져야 한다. 예를 들어, 조사와 어미의 결합형을 처리하기 위한 기준, 선어말 어미의 처리 기준, 복합명사의 수록 여부, '3·1절'이나 영문자와 아스키 문자의 포함 여부 등이 이에 속한다[1]. 즉, 형태소 분석의 기초 작업인 어휘사전은 형태소 분석 알고리즘과도 밀접한 관련이 있다.

어휘사전의 구성과 함께 알고리즘을 선택하는데 형태소 분석 알고리즘은 언어 독립적인 모델과 언어 종속적인 모델이 있다[2]. 언어 독립적인 모델은 형태론적 변형 문제를 중심으로 구현되므로 특정 언어의 특성과 관련된 문제 해결 부분이 취약하며, 언어 종속적인 모델은 타 언어에 적용하기는 어렵지만 그 언어의 특성에 적합한 결과를 생성하는데 적합하다. 특히, 한국어 어휘사전은 모든 형태소를 수록할 수 없기 때문에 사전에 수록되지 않은 것은 미등록어 추정 과정이 필요하고 미등록어 추정은 분석 결과에 미치는 영향이 매우 크다.

또한 형태소간의 결합 제약 및 축약 현상, 신문 기사와 같이 띄어쓰기를 무시하는 오류, 대화체 음성 언어에서 대화체에서만 사용되는 어휘 특성과 발화 오류 등 입력 유형이 다양하다. 따라서 한국어 형태소 분석기는 입력 유형의 다양성과 응용 분야의 적응성이 뛰어난 언어 종속적인 모델이 적합하며 한국어 형태소 분석에 필요한 정보들을 유용하게 활용할 수 있어야 한다.

한국어 형태소 분석기의 효율 및 성능에 미치는 요인들은 매우 많으나 본 논문에서는 한국어 단어의 특성과 관련된 것 중에서 어휘사전과 단어의 특성, 형태소 분석 결과의 특성 등 살펴 보고 한국어의 특성이 비교적 잘 반영되어 있는 형태소 분석기인 HAM(Hangul Analysis Module)[1]의 성능을 평가한다.

2. 형태소 분석과 어휘사전

형태소 분석 알고리즘은 형태소 사전에 의존하며, two-level 모델에서는 입력 문자열과 사전을 일치시키면서 two-level 규칙에 따라 음운 현상을 처리하기 때문에 trie 구조로 사전을 구성하여야 한다[3,4,5]. 한국어의 경우에도 분석 알고리즘에 따라 불규칙 활용 어간을 사전으로 처리하는 방법과 규칙으로 처리하는 방법이 있기 때문에 방법론에 따라 어휘사전의 구성이 달라진다. 더욱이 어휘사전의 구조와 access mechanism은 형태소 분석기의 처

1) version 1.5와 2.0이 있는데 version 2.0 library는 ham.hansung.ac.kr에서 anonymous ftp로 down받을 수 있다. 이 proceeding의 "한국어 형태소 분석을 위한 단어 유형 분류와 자료구조"(강승식) 참조.

리 속도에 미치는 영향이 매우 크기 때문에 어휘사전은 형태소 분석기의 성능 및 효율에 미치는 영향이 가장 크다.

2.1 어휘형태소 사전

국어사전은 어휘를 수록할 때 어휘형태소와 문법형태소를 하나의 사전으로 구성한다. 그러나 형태소 분석시에 어휘형태소와 문법형태소의 역할이 다르기 때문에 어휘형태소 사전과 문법형태소 사전을 별개로 구축한다. 어휘형태소 사전에 관한 논의는 주로 access 속도와 관련된 사전의 구조에 집중되어 왔다[6,7]. 이 때 어휘형태소의 출현빈도는 사전의 구조를 결정하는데 중요한 요소가 된다. 즉, 계층적으로 구축함으로써 access 속도를 매우 향상시킬 수 있다. 그러나 한국어 형태소에 대한 출현 빈도는 계량화되어 있지 않으므로 HAM에서는 형태소의 길이에 따라 사전을 논리적으로 분할하는 방법과 음절별로 가변 길이 record를 구성하는 방식을 혼합함으로써 access 속도를 향상시키는 방법을 취하고 있다.

HAM의 음절 record 방식은 한글 단어의 평균 음절수는 3.1 음절, 어근의 평균 음절수는 1.9 음절이라는 특성과 2 음절 어근이 가장 많이 출현한다는 특성을 이용한 것이다[8]. 이 방식은 서울대학교에서 개발된 KMA(Korean Morphological Analyzer)에서 사용된 block indexing 기법보다 access 성능이 매우 뛰어나지만, 오토마타로 구현하는 방식에 비해서는 성능이 떨어진다[6,8]. 그러나 오토마타 방식의 사전 구조가 주기억장치를 사용하는 방법을 취하고 있는데 비해 음절 record 방식은 하드디스크를 사용하는 방법을 취하고 있으므로 주기억공간을 적게 차지하는 장점이 있다.

표 1. 10만 단어 사전의 어휘수

품사 \ 어휘수	10만 어휘 사전		
명사, 대명사, 수사	1 음절	691	84,647 (81.63%)
	2 음절	41,907	
	3 음절	30,807	
	4+음절	11,242	
동사, 형용사, 보조용언	1 음절	360	11,139 (10.74%)
	2 음절	2,291	
	3 음절	4,583	
	4+음절	3,905	
부사, 관형사, 감탄사	부 사	6,757	7,906 (7.63%)
	관형사	755	
	감탄사	394	
총 어휘수	103,692(100%)		

어휘형태소 사전의 어휘수는 형태소 분석기의 성능에 직접 혹은 간접적인 영향을 미치고 있다. 일반적으로 많이 사용되고 있는 어휘형태소 사전의 어휘수를 품사별로 분류하면 표 1, 표 2와 같다. 표 1의 '10만 어휘 사전'은 일반적인 국어사전에 수록된 표제어 중에서 고어를 제외한 어휘형태소 10만 여개를 수록한 것이고, 표 2의 '6만 어휘 사전'은 국민학생용 국어 사전 등 소규모 사전을 중심으로 일상 생활에서 자주 사용되는 어휘들을 선별하여 수록한 것이다[9,10]. 이 때 명사 '먹', 동사 '먹-'과 같이 한 어휘가 두 가지 이상의 품사로 사용되는 것은 각각 따로 계산하였으며, 동사 '가-'와 보조용언 '가-'와 같이 같은 그룹에 속한 것은 하나로 계산하였다. 표 1, 표 2에 의하면 용언이나 부사어 등 체언을 제외한 나머지 어휘수는 1~2만 개로 사전에 수록된 전체 어휘의 대부분을 체언이 차지하고 있음을 알 수 있다.

표 2. 6만 단어 사전의 어휘수

품사 \ 어휘수	6만 어휘 사전		
명사, 대명사, 수사	1 음절	599	51,117 (82.42%)
	2 음절	22,405	
	3 음절	23,894	
	4+음절	4,219	
동사, 형용사, 보조용언	1 음절	323	5,460 (8.80%)
	2 음절	1,420	
	3 음절	2,242	
	4+음절	1,475	
부사, 관형사, 감탄사	부 사	4,438	5,445 (8.78%)
	관형사	720	
	감탄사	287	
총 어휘수	62,022(100%)		

'10만 어휘 사전'의 기반이 된 일반 국어 사전은 사용자가 잘 모르는 어휘들에 대한 설명이 주된 목적이므로 일상생활에서 드물게 사용되는 어휘가 다수 포함되어 있다. 그런데 어휘형태소 사전의 용도는 단지 어휘형태소라고 추정되는 문자열이 어휘형태소인지를 확인하는 기능이다. 따라서 문법형태소에 비해 형태소 분석 결과에 미치는 영향이 크지 않으며 추정에 의한 분석 결과의 생성이 가능하다. 즉, 사전에 수록된 어휘의 수가 형태소 분석에 미치는 영향은 긍정적인 면과 부정적인 면이 공존한다. 많은 어휘가 수록될수록 그 어휘에 대한 형태소 분석은 정확해지지만, 일상생활에서 드물게 사용되는 어휘가 형태소 분석 결과에 부정적인 영향을 미치는 경우도 흔히 발생한다.

예를 들어, '하고'(어떠한 이유라는 의미)라는 명사가 사전에 수록되어 있으면 '하$_{동사}$+고$_{어미}$'로 사용된 경우에도 항상 '하고$_{명사}$'로 분석하는 모호성이 발생한다. 이와 같이 드물게 사용되면서 모호성을 유발하는 어휘는 사전에서 제외하는 것이 더 나은 결과를 얻을 수 있다. 어휘형태소의 빈도수 정보가 이용되면 어휘의 수가 많은 사전이 더 좋은 형태소 분석 결과를 생성할 수도 있지만, 빈도수 정보를 이용하지 않는 형태소 분석기는 '6만 어휘 사전'을 이용한 형태소 분석기가 더 좋은 결과를 생성하게 된다.

HAM은 표 2의 6만 어휘 사전을 사용하고 있는데 다양한 유형의 문서에 대한 실험 결과 10만 어휘 사전을 사용하는 것보다 분석 결과가 매우 우수함을 알 수 있었다. 영어의 경우에도 6만개 정도의 어휘 사전이 가장 우수하다는 실험 결과가 있으며, 특히 한국어의 분석률은 미등록어 처리 결과가 더 큰 변수가 되는 특성이 있다.

2.2 문법형태소 사전

문법형태소 분석 사전은 형태소 분리 및 단어의 유형을 인식하는 등 형태소 분석에 미치는 영향이 매우 크다. 예를 들어, 어떤 문법형태소가 사전에 수록되지 않았다면

그 문법형태소가 사용된 단어에 대한 미등록 문법형태소를 추정하기가 어려우며 틀린 분석 결과를 생성하게 된다. 따라서 대화체 어휘에 대한 형태소 분석을 위해서는 대화체 문법형태소를 사전에 수록하여야 한다. 비표준 문법형태소가 포함된 문서에 대해서도 문법형태소를 사전에 수록하거나 별도의 처리 과정을 거쳐야 분석 오류를 막을 수 있다.

HAM에서 사용된 문법형태소 사전의 어휘수는 조사와 조사, 어말어미와 어말어미 등 결합형 문법형태소를 포함할 때 표 3과 같다. 표 3은 문법형태소의 음절수에 따라 조사와 어미의 수를 계산한 것으로 문어체에서 사용되는 어말어미와 조사를 수록하고 있다. 이 통계에서 대부분의 구어체 문법형태소와 드물게 사용되는 것은 포함되지 않았다. 또한 조사의 변이체 중에서 '께서', '께서는' 등과 같이 '께'로 시작되는 것은 그 수가 매우 많으나 '에게'로 시작되는 조사의 변이체로 간주하여 통계에서 제외하였다.

표 3. 조사/어미 사전의 어휘수

분류 / 음절수	조 사	어 미
1	29	48
2	89	197
3	138	294
4	106	176
5	52	50
6	14	23
7	2	0
합 계	430	788

조사와 어미는 각각의 특성에 따라 어근과 결합할 때 결합 제약 특성이 있으므로 이를 사전에 기술할 필요가 있다. 대부분의 어미는 동사와 형용사에 결합이 가능하지만 그렇지 않은 것도 있다. 어미의 유형에 따라 명령형 어미는 동사하고만 결합되며 '-다고'는 형용사와 결합 가능하지만 동사와는 결합할 수 없다. 그러나 '먹었다고', '먹겠다고'와 같이 선어말어미와 결합이 가능하다. 이와 같은 문법형태소의 결합 특성을 고려하여 문법형태소 사전에 기술되는 정보는 다음과 같다.

① 조사인지, 어말어미인지 혹은 둘 다 가능한지 여부
② 문어체, 구어체 여부
③ 어미의 경우에 동사, 형용사 결합 여부
④ 고빈도어, 저빈도어 여부

형태소 분석시에 문법형태소 사전은 형태소 분석기에 의하여 자주 참조된다. 그런데 조사와 어미는 고빈도어와 저빈도어로 명확하게 구별되므로 문법형태소 사전의 구조를 계층적으로 구성함으로써 사전 탐색 효율을 개선하거나 미등록어 추정시에도 조사의 고빈도, 저빈도 정보를 활용하여 미등록어 추정 효율을 높일 수 있다[11].

3. 한국어 단어의 특성

형태소 분석 결과는 정상적으로 형태소 분석에 성공한 경우(유형 1)와 형태소 분석에 실패하여 미등록어 추정에 의해 분석한 경우(유형 2)가 있다. 유형 1은 단어를 구성하고 있는 모든 형태소가 사전에 수록되어 있으므로 오분석 가능성이 거의 없으나, 유형 2는 미등록어가 포함되어 있으므로 오분석되었을 가능성이 있다. 유형 1은 단어에 모호성이 내포되어 있는지 여부에 따라 단일 분석 결과를 생성하는 경우, 2개 혹은 3개 이상의 분석 결과를 생성하는 경우로 구분된다.

표 4는 HAM version 1.5를 사용하여 실험한 것으로 어휘형태소의 품사 유형을 체언(noun), 용언(verb), 기타(adv, det, int)로 구별하고 있으며, 문법형태소의 유형은 조사(Josa), 어미(Eomi), 선어말어미(p-Eomi), 접미사(sfx)로 단순화시켜서 분석 결과를 생성하기 때문에 모호성이 발생하는 단어의 수가 많지 않다.

표 4. 한국어 형태소 분석 결과

분석결과 / 문서유형		논문요약	신문기사	문학작품	교과서
분석 성공어	1	105,464 (78.95%)	397,477 (68.82%)	209,341 (62.40%)	335,935 (73.65%)
	2	13,340 (9.99%)	59,861 (10.36%)	44,660 (13.31%)	76,630 (16.80%)
	3	4,120 (3.08%)	11,230 (1.95%)	9,247 (2.76%)	19,061 (4.17%)
	4	90 (0.07%)	6,764 (1.17%)	1,802 (0.54%)	4,283 (0.94%)
	5	0 (0.00%)	0 (0.00%)	3 (0.00%)	17 (0.00%)
추정어		10,570 (7.91%)	102,251 (17.70%)	70,452 (20.99%)	20,237 (4.44%)
총단어수		133,584 (100%)	577,583 (100%)	335,505 (100%)	456,118 (100%)

표 5. 한글 문서에서 체언/용언/부사어 출현 비율

어휘유형 / 문서유형	논문요약	신문기사	문학작품	교과서
체 언	87,540 (83.01%)	314,525 (79.13%)	131,167 (62.66%)	209,322 (62.31%)
용 언	14,704 (13.94%)	69,828 (17.57%)	63,195 (30.19%)	111,501 (33.19%)
부사어	2,742 (2.60%)	11,681 (2.94%)	12,712 (6.07%)	10,903 (3.25%)
기 타	478 (0.45%)	1,443 (0.36%)	2,267 (1.08%)	4,209 (1.25%)
총단어수	105,464 (100%)	397,477 (100%)	209,341 (100%)	335,935 (100%)

표 5는 표 4의 형태소 분석 결과 중에서 체언/용언/부사어/기타로 분류되는 어휘의 출현 비율을 계산한 것이다. 표 5에서 모호성이 있는 단어의 분석 결과는 어디에 포함시킬 것인지 판단하기가 어려우므로 모호성이 없는 단어, 즉 하나의 분석 결과를 생성하는 단어에 대해서만 계산하였다. 이 때 기타로 분류된 유형은 관형사, 감탄사이다. 각 유형의 상대적 출현 비율은 문서의 유형에 따라 차이가

있지만 전체 단어 중에서 체언의 비율이 62~83%를 차지하고 있으며, 용언은 14~33%이다.

표 4의 분석 결과에 의하여 형태소 분석 대상 어절의 대부분이 체언임을 알 수 있으며 미등록어의 대부분은 체언이다. 따라서 형태소 분석 과정은 단어의 비율에 따라 체언, 용언, 부사어의 순으로 분석하고 각 과정에서 체언 혹은 용언이 확실한 단어는 다른 가능성을 배제하는 방법을 사용할 수 있다.

표 6. 한글 문서에 나타난 단어의 평균 길이

문서유형 / 분석결과	논문요약	신문기사	문학작품	교과서	전 체
총단어수	14만	60만	34만	48만	156만
한 글 단어수	12만9천	49만3천	29만8천	42만8천	134만8천
음절수 / 단어	3.19	3.51	3.19	2.79	평균 3.17

4. 알고리즘의 효율 문제

일반적으로 알고리즘의 효율은 입력의 크기에 대한 단위 연산(unit operation)의 수로 계산되며, 최악 복잡도(worse-time complexity)와 평균 복잡도(average-time complexity)에 의하여 평가된다. 형태소 분석 알고리즘의 경우에도 입력 문자열의 길이(word length)에 대한 비교 횟수에 의해 알고리즘의 효율이 계산된다[12,13,14]. 그런데 표 6과 같이 몇 가지 유형의 한글 문서에 출현한 156만 단어에 대한 실험 결과 한글 단어의 길이는 평균 3.17 음절이고, 그 중에서도 2~4 음절어가 전체의 70~80%를 차지하고 있으며 10음절 이상인 단어는 거의 드물다(표 7). 즉, 한국어 형태소 분석의 입력 string은 특정 음절수로 이루어진 단어가 대부분이므로 단어의 길이 n을 기준으로 하는 최악 복잡도 O(n)에 의한 알고리즘의 평가는 적합하지 않다.

표 7. 한글 문서에 나타난 단어의 길이

문서유형 / 음절수	논문요약	신문기사	문학작품	교과서
1 음절	8432(6.54)	23382(4.74)	16482(5.53)	41581(9.71)
2 음절	31846(24.69)	112343(22.78)	74452(24.95)	142604(33.30)
3 음절	42508(32.95)	143679(29.14)	109538(36.72)	146396(34.19)
4 음절	26667(20.67)	104308(21.15)	57044(19.12)	67363(15.73)
5 음절	14368(11.14)	62599(12.69)	25531(8.56)	22773(5.32)
6 음절	3752(2.91)	24544(4.98)	8866(2.97)	5312(1.24)
7 음절	1006(0.78)	13095(2.66)	3802(1.27)	1877(0.44)
8 음절	302(0.23)	5274(1.07)	1518(0.51)	241(0.06)
9 음절	78(0.06)	2320(0.47)	599(0.20)	25(0.01)
10+음절	32(0.03)	1573(0.20)	518(0.17)	7(0.00)
총	128991(100%)	493123(100%)	298350(100%)	428179(100%)

한국어 형태소 분석은 단어의 음절 경계에서 형태소 분리가 가능한지를 검사하기 위해 형태론적 변형과 함께 사전 탐색이 빈번하게 발생한다. 그리고 한국어 단어는 길이에 대한 편차가 크지 않으므로 알고리즘의 효율은 '단어의 음절 경계'에서 형태소 분리가 가능한지를 검사하는 평균 복잡도가 알고리즘의 효율에 미치는 영향이 가장 크다. 한국어 형태소 분석 알고리즘은 형태소 분리를 위한 형태론적 변형과 사전 탐색으로 구성되므로 형태소 분리를 위한 비교 연산과 사전 탐색 연산이 알고리즘의 복잡도를 계산하는 요소가 된다. 특히 사전 탐색 연산은 알고리즘의 효율에 미치는 영향이 매우 크다.

분석 후보를 생성한 후에 옳은 분석 결과를 선택하는 형태소 분석 알고리즘은 분석 후보를 생성하기 위해 필요한 비교 연산과 생성된 후보에 대한 사전 탐색 횟수가 알고리즘의 효율을 좌우한다. Two-level 모델에서는 입력 문자열을 trie 구조 사전과 일치시키면서 분석을 행하므로 음운 현상을 처리하기 위해 사전 탐색시에 발생하는 backtracking 연산이 효율을 좌우한다. 즉, 형태소 분석의 효율은 사전 탐색에 크게 의존하고 있다.

사전 탐색 연산이 형태소 분석에 미치는 영향을 측정하기 위하여 HAM version 1.5에서 형태소 분석에 소요되는 시간을 사전 탐색 시간과 기타 연산 시간으로 구분하여 측정하였다. 실험 방법으로는 정상적으로 형태소 분석을 한 경우와 단지 어휘형태소 사전만 탐색하지 않은 경우에 걸리는 시간을 각각 측정하였다. 즉, 문법형태소 사전을 이용하여 형태소 분리 및 형태론적 변형 과정을 거쳐 생성된 분석 후보에 대하여 어휘형태소 사전을 탐색한 경우와 그렇지 않은 경우로 나누어 실험하였다. 그 결과, 어휘형태소 사전을 탐색하지 않았을 때 처리 속도가 4.8배로 증가하였다. 실험 결과에 의하면 전처리, 형태소 분리, 형태론적 변형 등 형태소 분석을 위한 제반 처리 비용에 대한 어휘형태소 사전의 탐색 비용이 3.8배나 되며, 형태소 분석에 소요되는 시간의 약 80%가 사전을 탐색하는데 소요됨을 알 수 있다.

(사전 탐색 시간) : (기타 연산 시간) = 3.8 : 1

이 실험은 사전 탐색 알고리즘에 의존적이며 형태소 분석 기법에 따라 편차가 있을 수 있으므로 객관적인 결과는 아니다. 실험에 사용된 사전 탐색 알고리즘은 블록 인덱스 탐색에 의하여 특정 음절로 시작하는 어휘형태소에 대한 블록을 찾고 블록내에서는 이진 탐색 알고리즘을 적용한 음절 record 방식이다. 그러나 two-level 모델의 경우에도 사전 탐색 시간이 분석 시간의 대부분을 차지하고 있으므로 형태소 분석에서 알고리즘의 효율에 가장 큰 영향을 미치는 것은 사전 탐색이라 할 수 있다.

5. 형태소 분석기 구현시 고려 사항

형태소 분석과 철자 검사는 단어 형성 규칙에 따라 형태소 사전과 형태론적 변형 규칙을 기반으로 구현된다. 한국어 형태소 분석의 특성에 따라 효율적인 범용 형태소 분석기를 구현할 때 고려되어야 할 점은 다음과 같다.

첫째, 단어 형성 규칙에 어긋나는 오류어는 분석 실패 후 미등록어 추정을 한다. 옳은 단어만을 처리하는 형태소 분석기는 입력 오류를 고려하지 않아도 되지만 철자 검사나 자동 색인 등 대부분의 응용 시스템에서는 오류어 또

는 미등록 어휘형태소를 처리해야 하기 때문이다. 입력 단어가 단어 형성 규칙에 어긋나는지를 검사하는 과정이 철자 검사이며 그 과정에서 단어 형성 규칙에 맞는 분석 결과를 생성하는 것이 형태소 분석이다. 따라서 형태소 분석 알고리즘은 단어 형성 규칙에 맞는 단어는 인식하지만 그렇지 않은 단어는 인식하지 못하게 하는 역할을 한다.

둘째, 형태소 사전은 형태소 분석에 반드시 필요한 사전과 추가 정보가 수록된 사전으로 구분하여 구성한다. 형태소 사전에 반드시 수록되어야 하는 어휘형태소는 '감기는'에서 명사 '감기', 동사 '감다'와 같이 모호성을 유발하는 형태소이다. 한국어는 형태소를 분리하고 단어의 유형을 인식할 때 분석 후보가 한 가지만 생성되는 경우에 더 이상 사전을 참조할 필요가 없다. 또한 형태소 분석이나 철자 검사에 사용되는 어휘형태소 사전의 어휘수는 많을수록 더 좋은 결과를 생성하지는 않는다. 어휘형태소 사전의 크기가 클수록 명사가 차지하는 비율이 높고 이 때 상대적으로 빈도가 매우 낮은 어휘형태소가 많이 수록되므로 이로 인한 오분석률이 높아진다. 즉, '하고', '법의', '망의', '위해' 등과 같은 저빈도 명사가 수록되면 특정 유형의 데이터에 대하여 옳은 분석 결과를 생성할 수 있으나 일반적인 데이터에 대해서는 모호성이 증가할 뿐 아니라 철자 검사에서도 '하고를'과 같은 오류어를 발견하지 못하게 된다. 따라서 사전의 크기가 커지면 각 어휘형태소에 대한 빈도 정보를 수록하거나 또는 빈도수에 따라 일상적인 어휘 사전과 드물게 사용되는 어휘 사전, 분야별 어휘 사전 등 계층적으로 구성한다.

셋째, 형태소 분석의 효율은 입력 단어의 길이가 아니라 단어를 분석하기 위하여 시도되는 평균 분석 시도 횟수로 계산된다. 한국어 형태소 분석은 단어 구성 전이도와 형태론적 변형 규칙에 따라 조합 가능한 탐색 공간(search space)에서 옳은 경로(path)를 찾아 가는 문제이다. 즉, 단어의 길이가 길다고 해서 반드시 탐색 공간의 크기가 커지는 것은 아니며 입력 문자열을 구성하는 자모 혹은 음절의 유형에 따라 탐색 공간의 크기가 결정된다. Two-level 모델에서는 입력 문자열을 사전과 일치시키므로 긴 단어에 대한 사전 일치 시간이 길어지지만 이 경우에도 형태론적 변형 가능성이 많은 문자열은 빈번한 backtracking을 유발하게 되고 실제로 backtracking 횟수가 단어의 길이에 비해 알고리즘의 효율에 미치는 영향이 더 크다.

넷째, 형태소 분석 알고리즘은 탐색 공간을 최소화하고 확장성을 최대로 하는 방법을 취한다. 형태소 분석기의 탐색 공간은 입력 문자열의 특성에 따라 결정되며, 입력 문자열에 대하여 모든 분석 가능성을 검사하여 가능성이 있는 후보를 찾은 후에 그 후보에 대해 분석 결과를 생성한다. 분석 가능성이 많은 문자열이 포함된 단어는 모호성을 유발할 가능성이 높으며 또한 탐색 공간이 커진다. 따라서 탐색 공간을 최소화하려면 분석 가능성 검사 조건을 강화해야 한다. 그런데 제약 조건을 강화하면 확장성(flexibility)이 저하되어 미등록어 추정에 방해가 될 수 있으므로 확장성을 저하시키지 않는 범위내에서 제약 조건을 강화하여야 한다.

다섯째, 형태소 분석 결과에 가중치를 부여한다. 여러 개의 분석 결과 중에 분석 성공한 것이 반드시 옳은 것은 아니며 "유가와 이론"의 "유가와"처럼 실제로는 미등록어가 포함된 단어가 전혀 엉뚱한 분석 성공 결과를 생성할 수도 있다. 분석 결과에 대한 가중치는 분석 성공한 결과에 대하여 가중치를 높이고 분석 실패한 분석 후보는 가중치를 낮추며, 분석 성공한 결과에 대해서도 단어를 구성하고 있는 형태소들의 빈도 정보와 단어 유형에 따라 가중치를 부여할 수 있다. 분석 실패한 후보에 대해서도 가중치에 의하여 분석 결과를 생성하는 이유는 '미테랑'을 '미테'$_{\text{미등록명사}}$+'-랑'$_{\text{조사}}$ 또는 '미테랑'$_{\text{미등록명사}}$로 분석이 가능하며, '홍부가'의 경우에도 '홍부'$_{\text{명사}}$+'-가'$_{\text{조사}}$ 또는 '홍부가'$_{\text{미등록명사}}$로 분석될 수 있기 때문이다.

6. 형태소 분석기를 이용한 철자 검사

철자 검사는 입력 문장에 오류어가 포함되어 있는지를 검사하는 것으로 오류어의 발견, 오류 형태소의 교정, 그리고 문법 검사를 위하여 형태소 분석 기능이 필요하다. 철자 검사에서 형태소 분석의 기능은 '철자 검사를 위한 형태소 분석'과 '형태소 분석을 이용한 철자 검사'의 두 가지 유형이 있다[15]. 철자 검사를 위한 형태소 분석에서는 오류어를 발견하거나 교정하는 것이 목적이므로 형태소 분석을 기반으로 하여 오류어 발견을 위한 알고리즘 및 어휘형태소 사전이 구현된다. 예를 들어, 자주 발생하는 오류어들에 대하여 그 특징을 규칙 형태로 기술하거나 사전에 수록함으로써 형태소 분석 단계에서 처리하기도 한다.

철자 검사를 위한 형태소 분석은 오류어의 발견 및 교정을 중심으로 기술되므로 일반적인 형태소 분석과는 차이가 있으며 현재 문서편집기에 내장된 대부분의 한글 철자 검사기는 이러한 유형에 속한다. 따라서 철자 검사를 목적으로 구현된 형태소 분석 기능은 다른 응용 시스템에 적용하기가 어렵다. 형태소 분석기를 철자 검사에 이용하면 형태소 분석 결과를 철자 오류 교정, 문법 검사(grammar check), 주제어 검색, 문서 요약 등 문서 편집기에 추가될 기능에서 공유할 수 있는 장점이 있다. 형태소 분석기를 철자 검사에 활용하기 위해서는 오류어 처리 기능이 강화되어야 하며 오류어 처리를 고려한 형태소 분석 기능은 다음과 같다.

철자 검사를 위한 형태소 분석은 제약 조건이 철저하게 검사되어야 한다. 이에 비해 일반적인 형태소 분석에서는 입력 단어에 오류가 없다고 가정되며 오류어에 대한 형태소 분석은 고려할 필요가 없다. 예를 들어, 한국어 이해 시스템과 같이 입력 단어에 오류가 없다고 전제되는 형태소 분석에서는 조사 '-를'을 분리할 때 어휘형태소가 무종성 체언인지를 검사하지 않아도 된다. 또한 범용 형태소 분석기는 용언의 불규칙 활용에서 '아름다와서'와 같이 잘못된 어미 활용에 대한 제약을 검사하지 않아도 된다.

입력 오류를 고려하지 않은 형태소 분석기를 철자 검사기에서 사용하면 철자 오류를 발견할 수 없는 경우가 발생한다. 입력 단어가 오류어인지 아닌지를 판단하기 위해서 형태소 분석 결과를 이용하는데 형태소 분석 결과는 단어를 구성하고 있는 각 형태소에 철자 오류가 포함되어 있을 가능성을 고려해야 하고 이웃한 형태소끼리의 결합 가능 조건뿐 아니라 결합 불가능 조건까지도 검사해야 한다.

철자 검사는 형태소 분석기에서 중요시되는 모호성 문제가 상대적으로 덜 중요하다는 점에서 형태소 분석기의 성능을 판단하기 어려운 점이 있다. 그러나 빠른 처리 속도를 요구하기 때문에 형태소 분석기의 처리 속도를 평가하기에 적합하다. 형태소 분석기를 철자 검사를 위한 용도로 사용할 때는 형태소 분석기가 틀맞 오류와 맞틀 오류를 발생하지 않도록 설계되어야 한다. 또한 음절 특성을

철자 검사시에 활용할 수도 있다.

철자 오류어에 대한 교정 후보를 제시하는 기능(spell aid)은 오류의 범위를 최소화하기 위하여 형태소 분리가 요구되므로 철자 검사기는 형태소 분석기를 기반으로 구현되는 것이 효율적이다. 영어에서는 철자 검사에서 문장의 구조적 오류를 발견하는 문법 검사 기능이 제공되고 있으나, 한글 철자 검사기는 아직 문법 검사 기능이 제공되지 않고 있다. 형태소 분석을 이용한 철자 검사기에서는 형태소 분석 결과에 구문 분석 결과를 활용함으로써 문법 검사 기능을 추가하기가 용이하다.

7. 실험 결과 및 성능 평가

HAM은 형태소 분석 및 철자 검사, 자동 색인 등이 가능하도록 구성된 단어 구성 전이도와 음절 특성을 이용하여 단어의 유형을 인식하는 방법을 사용한다. 어휘형태소 사전은 현대 국어에서 주로 사용되는 6만여 개의 어휘를 수록하고 있으며, 문법형태소 사전은 조사와 어미의 결합형을 사전에 수록하는 방식을 사용하고 있다. 형태소의 품사 체계는 체언, 용언, 기타(부사, 관형사, 감탄사)로 분류하여 모호성의 수를 줄이고 분석 결과를 단순화시키는 방법을 취하고 있다. 이 형태소 분석기는 범용 형태소 분석기능을 지향하고 있으므로 분석 결과의 생성뿐 아니라 실행 option에 따라 철자 검사 기능 또는 자동 색인 기능으로 작동하기도 한다. 여기서는 형태소 분석 결과에 대한 분석률과 철자 검사기로 사용했을 때의 실험 결과를 기술한다.

7.1 형태소 분석 기능에 대한 실험

형태소 분석기의 분석 결과를 실험하기 위한 데이터 집합은 여러 가지 유형의 데이터에서 무작위로 약 3,000 어절씩 추출한 것으로 각 데이터 집합의 크기는 표 8과 같다.

표 8. 실험 데이터의 크기

국민학교 교과서	신문기사	전산학 논문	문학작품	이규태 칼럼	합 계
3,221	3,177	3,025	3,069	3,023	15,515

형태소 분석기가 출력한 결과를 수작업으로 확인하여 분석 성공한 어절을 계산하였다. 실험 데이터에서 철자 오류와 띄어쓰기 오류 어절은 수작업으로 교정하였으나 교정되지 않은 어절은 분석에 성공한 것으로 계산하였으며, 인용 부호나 괄호 뒤에 오는 조사는 본 형태소 분석기에서 분석하지 않고 있으므로 분석 대상에서 제외하였다. '저녁때'와 같이 접미사를 따로 분리하지 못한 것과 분석 실패하여 미등록어 추정에 의하여 단일어로 간주한 것도 분석 성공한 것으로 계산하였다. 형태소 분석 결과에 모호성이 발생하여 두 가지 이상의 결과를 생성했을 때 그 중에서 옳은 것이 있으면 분석 성공한 것으로 간주하였다. 각 데이터 집합에 대한 분석 결과는 표 9와 같다.

표 9의 결과는 형태소 분석 결과가 2 개 이상인 경우에 하나라도 옳은 것이 있으면 분석 성공으로 간주한 것으로 HAM version 1.5의 평균 분석률은 99.46%이다. 그러나 분석결과 중에 옳은 것이 있다 하더라도 잘못 분석된 결과가 있으면 분석 실패로 간주했을 때의 평균 분석률은 98.9%로 나타났다. 이와 같이 오분석 결과가 포함된 것은 미등록어 추정시에 옳은 결과를 생성하기도 하지만 옳지 않은 분석 후보를 생성하기 때문에 발생한다. 형태소 분석에 실패한 어절의 유형을 살펴 보면 다음과 같다.

표 9. 한국어 형태소 분석 결과

국민학교 교과서	신문기사	전산학 논 문	문학작품	이규태 칼 럼	합 계
99.44%	98.90%	99.87%	99.71%	99.40%	99.46%

첫번째 유형은 형태소 분석기가 처리하지 못하는 단어 유형으로 '즐거워하였습니다'와 같이 'ㅂ'-불규칙 용언 뒤에 '-하다'가 결합된 유형과 '창식이도'처럼 인명 뒤에 조음소 '이'가 삽입되는 어절, '사용되어온'이나 '전해달라고'와 같이 두 개의 용언이 결합된 것, '순리아닌'이나 '조건지우게'와 같이 조사가 탈락하면서 다음 용언과 결합된 유형, '됐음직하다'처럼 문법형태소가 미등록어인 경우, '교훈이랄'과 같은 축약어 등이 있다. 두번째 유형은 영문자나 아라비아 숫자가 포함된 것으로 '6백20달러선에서'나 '1월1일부터'와 같이 한글과 숫자가 복합된 일부 유형은 문법형태소 분리를 하지 않았기 때문에 분석에 실패한 것이다[2]. 마지막 유형으로는 미등록어 추정이 잘못된 경우로 '온나라', '서양말로' 등이 있다.

7.2 철자 검사 기능에 대한 실험

철자 검사 기능에 대한 실험을 위하여 두 가지 형태의 실험 데이터를 구성하였다. 첫번째 데이터 집합은 여러 사람이 입력한 한국어 사전의 표제어 중에서 철자 오류가 발생한 표제어 559개를 선정한 것이고, 두번째 유형은 시스템 공학 센터에서 실험용으로 만든 오류 문서 중에서 오류어라고 간주되는 어절 303개를 추출하였다. 실험에 사용된 철자 검사기는 본 논문에서 구현된 형태소 분석을 이용한 철자 검사기와 한글 2.5와 MS word 6.0의 철자 검사기이다.

표 10. 철자 검사기의 실험 결과

오류어 집합 / 맞춤법검사기	오류어 집합 1 (559 어절)	오류어 집합 2 (303어절)
HAM 1.5	423 (75.67%)	289 (95.38%)
한글 2.5	440 (78.71%)	281 (92.74%)
MS word 6.0	424 (75.85%)	263 (86.80%)

각 철자 검사기가 각 오류어 집합에 대하여 발견한 오류어의 수 및 퍼센트는 표 10과 같다. 철자 검사기의 성능은 재현율(recall ratio)과 정확률(precision ratio), 그리고 실행 속도에 의하여 평가되어야 한다. 또한 철자 교정 기능까지도 비교 평가해야 한다. 그런데 표 10에서는 재현율

2) HAM version 2.0은 숫자나 영문자가 포함된 단어에 대한 형태소 분석 기능이 보완되었으므로 본 논문의 실험 결과보다 분석률이 더 높다.

에 대해서만 비교한 것이며 데이터의 유형 및 오류어의 수가 많지 않기 때문에 객관적인 비교 평가라고 할 수는 없다. 그러나 표 10에서 알 수 있듯이 본 논문에서 형태소 분석기를 이용한 철자 검사기가 상용 시스템에 비하여 성능이 떨어지지 않음을 알 수 있다. 기존의 철자 검사기들은 상용화하기 위하여 수 개월 동안 테스트 및 tuning 작업을 거친 것이므로, 본 논문에서 구현된 형태소 분석기도 테스트 및 tuning 작업을 거친다면 기존의 철자 검사기보다 성능이 더 우수할 것이다.

7.3 형태소 분석기의 처리속도

KMA의 형태소 분석 속도는 IBM-PC 386 DOS환경에서 약 35 단어/초이다[8]. 이에 비해 HAM은 PC pentium 120MHz linux 2.0 환경에서는 약 1,000 단어/초이고, PC pentium 100MHz DOS 6.0 환경에서 약 300 단어/초이다.

8. 결 론

형태소 분석기는 형태소 분석이 요구되는 다양한 응용분야에서 활용될 수 있도록 구현되어야 한다. 이러한 형태소 분석기를 설계 및 구현할 때 형태소 분석에 영향을 미치는 요인들로 어휘사전과 한국어 단어의 특성을 살펴 보았으며, 또한 형태소 분석 기능과 철자 검사 기능을 충족시키기 위한 원론적인 요건들을 제시하였다. HAM은 범용 한국어 형태소 분석기를 지향하여 설계 및 구현되어 형태소 분석 기능뿐만 아니라 정보 검색 시스템에서 자동 색인 기능으로 사용되고 있으며 간단한 철자 검사 기능과 교정 후보 제시 기능이 있다.

본 논문에서는 형태소 분석기가 다양한 기능으로 활용될 수 있음을 보이기 위하여 형태소 분석 기능과 철자 검사 기능에 대하여 그 성능을 실험하였다. 실험 결과, 형태소 분석 기능으로 평균 99.46%의 분석률로 서울대학교 KMA의 99.36%보다 더 좋음을 알 수 있다. 더욱이 KMA는 10만 단어 사전을 사용함으로써 6만 단어를 사용하는 HAM보다 모호성이 더 많기 때문에 전반적으로 평가할 때 HAM의 성능이 더 우수하다.

HAM의 철자 검사 기능은 상용화되어 있는 두 개의 철자 검사기와 비슷한 성능을 보이고 있는데 다른 철자 검사기로서 설계되었기 때문에 다른 용도로 사용이 불가능하다는 단점이 있다. 즉, HAM은 문서 관리 시스템이나 그룹웨어, 인터넷 검색 시스템 등과 같이 철자 검사 및 자동 색인, 기계 번역, 음성 인식 등 형태소 분석을 기반으로 하는 소프트웨어가 통합된 시스템에 매우 적합하다.

특히 형태소 분석 속도가 시스템 환경에 따라 차이가 있으나 초당 300~1,000 단어로 철자 검사나 디지틀 도서관의 자동 색인 등 빠른 처리 속도가 필요한 경우에 더욱 적합하다.

참고문헌

[1] 강승식, 김영택 "사전 정보에 기반한 효율적인 한국어 형태소 분석기의 설계 및 구현", 정보과학회 춘계 학술발표 논문집, 18권 1호, pp.529-532, 1991.

[2] 강승식, "한국어의 형태론적 특성과 형태소 분석 기법", 정보과학회지, 12권 8호, pp.47-59, 1994.

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

[4] H.C. Kwon, L. Karttunen, "Incremental Construction of a Lexical Transducer for Korean," Proceedings of the 15th International Conference on Computational Linguistics, Vol.2, pp.1262-1266, 1994.

[5] D.B. Kim, S.J. Lee, K.S. Choi, G.C. Kim, "A Two-level Morphological Analysis of Korean," Proceedings of the 15-th International Conference on Computational Linguistics, Vol.1, pp.535-539, 1994.

[6] 백대호, 이 호, 임해창,, "Finite State Transducer를 이용한 한국어 전자 사전의 구조", 제7회 한글 및 한국어 정보처리 학술발표 논문집, pp.181-187, 1995.

[7] 이승선, 송주원, 조완섭, 황규영, 최기선, "Compact TRIE Index: 한국어 전자 사전을 위한 데이타베이스 색인 구조", 정보과학회 논문지, 22권 1호, pp.3-12, 1995.

[8] 강승식, *음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석*, 서울대학교 컴퓨터공학과 박사학위 논문, 1993년 2월.

[9] 금성출판사 사서부, 뉴에이스 국어 사전, 금성출판사, 1989.

[10] 민중서림 편집국, 국민학교 민중 새국어사전, 민중서림, 1992.

[11] 강승식, "상대적 출현빈도를 이용한 조사/어미 사전의 구성", 제7회 한글 및 한국어 정보처리 학술발표 논문집, pp.188-194, 1995.

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

[13] E. Barton, "Computational Complexity in Two-Level Morphology," 24th Annual Meeting of the Association for Computational Linguistics, 1986.

[14] 이은철, 이종혁, "계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현", 제4회 한글 및 한국어 정보처리 학술발표 논문집, pp.95-104, 1992.

[15] 권혁철, 채영숙, 김재원, 김민정, "한국어 철자 검색을 위한 형태소 분석 기법", 국어정보학회 학술발표 논문집, pp.179-186, 1991.