어절 분석 기반 형태소 분석 시스템 개발에 관한 연구

A Study on the Development of a Practical Morphological Analysis System Based on Word Analysis

조현양(Hyun-Yang Cho)*, 최성필(Sung-Pil Choi)**, 최재황(Jae-Hwang Choi)***

초 록

본 연구에서는 정보검색시스템의 성능향상을 위하여 기존에 연구되었던 다양한 어절 분석 기법들을 바탕으로 어절 분석 속도의 최대화, 형태소 분석기의 모듈화 및 구조화 그리고 형태소의 정확한 분석을 위한 한국어 어절 분석 시스템을 개발하였다. 본 연구에서 개발된 시스템은 어절 분석 속도를 높일 수 있는 최적의 알고리즘을 구현하였으며, 모듈화된 하부 시스템의 유기적이고 효율적인 결합에 중점을 두고 각 모듈별 성능 및 속도 검증이 가능하도록 하였다. 또한, 재귀적 복합명사 분석을 탈피하여 시스템 부하를 줄이고 다층적 수사 패턴 인식에 기반한 수사 형태소 분석 시스템을 개발하였다. 개발된 어절 분석 시스템을 이용하여 색인 시스템을 구성하고 이를 기반으로 실험을 하였다.

ABSTRACTS

The purpose of this study is to develop a Korean word analysis system, which can improve performance of IRS, based on various methods of word analysis. In this study we focused on maximizing the speed of Korean word analysis, modulizing each functional system and analyzing Korean morpheme precisely. The system, developed in this study, implemented optimal algorithm to increase the speed of word analysis and to verify speed and performance of each subsystem. In addition, the numeral analysis processing was achieved to reduce a system burden by avoiding recursive analysis of compound nouns, based on numeral pattern recognition.

키워드 : 형태소 분석기, 어절 분석, 형태소 분석
morphological analyzer, word analysis, morphological analysis

* 한국과학기술정보연구원 책임연구원 (hycho@kisti.re.kr)
** 한국과학기술정보연구원 연구원 (spchoi@kisti.re.kr)
*** 한국과학기술정보연구원 선임연구원 (findit@kisti.re.kr)

■ 논문 접수일: 2001년 5월 14일
■ 게재 확정일: 2001년 6월 8일

1 서 론

형태소 분석 기술은 자동 색인 시스템을 구성함에 있어서 가장 필수적이며, 근간이 되는 기술이다. 형태소 분석은 기계 번역이나 자연어 이해 시스템(natural language understanding system)을 구현하기 위하여 구문 분석과 의미 분석의 전 단계로 시작되었다. 최근에는 형태소 분석 성공률이 99% 이상의 실용화 수준에 이르고 있으나, 알고리즘의 효율성이나 복합명사 및 사전 미등록어 처리, 모호성 제거 문제, 단어의 사용 빈도에 따른 사전의 구성 등 개선의 여지가 남아 있다. 형태소 분석 알고리즘의 효율성은 입력 단어(어절)로부터 분석 후보를 생성하는데 필요한 비교 연산(comparison)의 수와 사전 검색 횟수에 의하여 결정된다. 비교 연산은 형태소의 분리와 불규칙 어절의 원형 복원 과정에서 주로 발생하고 사전 검색 횟수는 분석 후보의 수에 비례하는 것으로 나타났다(강승식 1993).

최근 들어 다양한 형태소 분석 시스템이 개발되고 있으며, 실제로 다양한 분야에 적용하기 위하여 독특한 자료구조와 알고리즘이 적용되고 있다. 자연어처리 시스템, 특히 한국어 처리 시스템의 가장 중요한 요소는 언어의 유연한 확장성이나 생성 현상, 혹은 신조어나 고유명사, 전문용어들에 대한 시스템의 유연성과 확장성이다. 형태소 분석 시스템은 우선, 시스템개발자 측면에서는 어절 분석 요소들이 전체 시스템의 성능에 직접적인 영향을 주기 때문에 좀더 쉽고 효율적인 방법으로 분석 시스템 자체를 변경하고 성능향상을 도모할 수 있어야 한다. 또한 시스템관리자 측면에서 형태소 분석 시스템은 사전 엔트리를 보다 효율적으로 관리하고 사용자 정의 사전에 대한 다양한 처리를 기반으로 새로운 언어 현상에 능동적으로 대처할 수 있어야 한다. 그러나 대부분의 기존 시스템들은 구조의 복잡성, 혹은 어절 생성 현상에 부적절한 알고리즘이나 자료구조로 인해 변경이나 업데이트가 불가능하다. 따라서 이러한 문제점들을 해결하기 위해서 본 연구에서는 한국어 어절 생성 규칙을 적용한 형태소 분석 자료구조 및 알고리즘을 개발하고, 이를 쉽게 변경하고 관리할 수 있는 시스템 구조로 구현하였으며, 분석 속도를 높이기 위한 다양한 최적화 알고리즘을 효과적으로 적용하였다.

본 연구의 목적은 정보검색시스템의 성능향상을 위하여 기존에 연구되었던 다양한 어절 분석 기법들을 바탕으로 자동 색인 시스템을 구성하는 형태소 분석 기술로써 어절분석 속도의 최대화, 형태소 분석기의 모듈화 및 구조화 그리고 형태소의 정확한 분석을 위한 한국어 어절 분석 오토마타를 효율적으로 분석할 수 있는 한국어 어절 분석 시스템을 개발하는 것이다.

2 형태소 분석기술의 특징

정보 검색 시스템의 한 부분인 자동 색인 시스템을 구성하는 형태소 분석 기술은 다음과 같은 특징을 가지고 있어야 한다. 첫째, 대용량 입력 문서를 빠른 시간 내에 정보 검색 시스템에 적재할 수 있도록 어절 분석 속도 최대화가 가능한 기술이 포함되어야 한다. 둘째, 전체 시스템 구조가 효율적으로 구성된 정보 검색 시

스템의 하부 엔진으로 포함되기 위해서는 형태소 분석 모듈 또한 자체적으로 모듈화와 구조화가 이루어져야 한다. 이는 대부분의 자연어 처리 시스템의 문제점 가운데 하나로써 복잡한 시스템 구조로 인한 시스템 관리나 응용의 어려움을 최소화하여 전체 정보 검색 시스템의 확장성을 도모한다는 차원에서 의미가 있다. 셋째, 형태소에 대한 정확한 분석을 위해서는 한국어 어절 분석 오토마타에 나타나는 모든 필요한 분석 아이템에 대해 정확한 분석 방법을 제시하고 이를 효율적으로 처리할 수 있어야 한다(강승식 1993). 예를 들어, 대부분의 시스템이 용언 분석 과정에서 선어말어미나 어말어미처리, 규칙, 불규칙처리 방법론에 대해서는 강조하는 반면, 보조용언이나 아/어 변이체 처리, 수사 처리 등에 대해서는 적절한 시스템 구현 방법이나 규칙을 기술하지 않고 있다. 이것은 전체 시스템의 성능에 크게 영향을 주지 않을 수도 있으나 시스템의 확장성이나 보다 정확한 어절 분석을 위한 요소 기능으로 매우 중요한 부분이다(채영숙 외 1998).

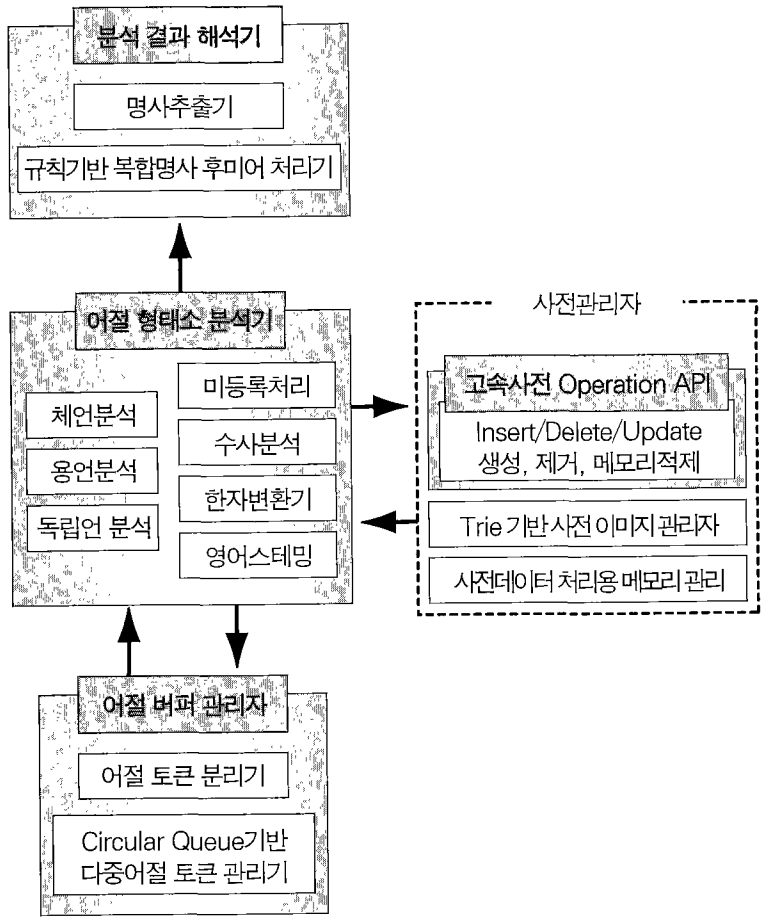
정보검색의 효율을 증대시키기 위하여 본 연구에서 개발될 시스템은 기존의 시스템과 비교하여 최소한 다음과 같은 특징을 포함하고 있다.

- 어절 분석 속도를 높이기 위하여 품사 사전의 구조화와 탐색 방법, 또한 이를 위한 다양한 접근 방법에 대한 평가를 통해 최적의 알고리즘이 구현되어야 한다.
- 전체적인 시스템 구조를 디자인함에 있어서 완벽하게 모듈화된 하부 시스템의 유기적이고 효율적인 결합에 중점을 두고 각 모듈별 성능 및 속도 검증이 가능하여야 한다.
- 대부분의 형태소 분석 시스템에서 적용하고 있는 재귀적(recursive) 복합명사 분석을 탈피하여 빈번한 재귀적 호출에 따른 시스템 부하를 줄이고 확장성을 도모할 수 있어야 한다.
- 다층적 수사 패턴 인식에 기반한 수사 형태소 분석 시스템을 개발하여 형태소 분석 시스템과 결합할 수 있어야 한다.
- 다어절 분석을 위한 기반 구조를 제공하기 위해서 원형 큐(Circular Queue) 기반 다어절 토큰 관리기를 개발하여 확장성을 극대화할 수 있어야 한다.

3 어절 분석 시스템의 구조

본 연구에서 개발하고자 하는 어절 분석 시스템의 전체적인 구조는 다음의 〈그림 1〉과 같다.

어절 버퍼 관리자는 입력된 어절 버퍼를 분리하여 각각의 어절 토큰을 생성하는 어절 토큰 분리기와 분리된 어절 토큰을 원형 큐에 저장하여 다어절 윈도우를 생성하는 다중어절 토큰 관리기로 구성된다. 다중어절 토큰 관리기는 현재 어절을 중심으로 앞뒤 각각 다섯 어절을 참조할 수 있는 API(Application Program Interface)를 제공하게 된다. 사전 관리자는 Tree 구조를 기반으로 시스템에서 사용하는 다양한 사전들을 저장하고 이를 고속으로 접근할 수 있는 다양한 API를 제공하며, 실제 어절 분석 모듈인 어절 형태소 분석기는 한국어 어절의 구성 형태에 따른 유한 오토마타의 각 상태 전의 구조로 이루어진다. 분석 결과 해석기는 어절에 대한 최종 형태소 분석 결과를 입력받아서 유효한 명



〈그림 1〉 시스템 구조

사를 추출함과 동시에 복합명사 분석 결과에 대한 기본적인 모호성 처리를 수행한다.

위에서 설명된 하부 시스템들은 각각의 상관관계를 설명하는 API로 밀결합 되어 있다. 이러한 밀결합이 전역변수나 기타 복잡한 설정으로 이루어진 것이 아니라 완벽한 상관관계 API로 구성되어 있으므로 모듈화에 따른 전체 시스템의 확장성 및 관리 효율성이 보장되도록 구성되었다.

4 어절 분석 시스템

4.1 어절 분석 사전의 구조와 사전 접근 API

가. 어절 분석 사전의 구조

분석 사전의 구성과 접근 방법은 형태소 분석 시스템의 전체 성능을 좌우할 수 있는 부분이다. 따라서 형태소 분석 시스템이 사용하는 품

〈표 1〉 분석 사전정보

정 보	품 사	정 보	품 사
1	부사	42	
2	지시 부사	43	
3	문장 접속 부사	44	
4	단어 접속 부사	45	
5	시간성 부사	46	
9		50	형용사
10	독립어	51	지시 형용사
11		52	동사
12		53	'ㅅ' 불규칙 동사
13		54	'ㅅ' 불규칙 형용사
14		55	'ㄷ' 불규칙 동사
15		56	'ㅂ' 불규칙 동사
16		57	'ㅂ' 불규칙 형용사
17		58	'ㅎ' 불규칙 형용사
18		59	'ㅎ' 불규칙 지시형용사
19		60	'우' 불규칙 형용사
20	관형사	61	'러' 불규칙 형용사
21	지시 관형사	62	'러' 불규칙 동사
22	수 관형사	63	'르' 불규칙 형용사
23		64	'르' 불규칙 동사
24		65	'으' 탈락 동사
25		66	'으' 탈락 형용사
26		67	'ㄹ' 탈락 동사
27		68	'ㄹ' 탈락 형용사
28		69	'하' 불규칙 형용사
29		70	'하' 불규칙 동사
30	불완전명사	71	'하' 불규칙 지시형용사
31	단위명사	72	'와' 축약 동사
32	보통명사	73	'이' 축약 동사
33	동작성 보통명사	74	'이' 축약 형용사
34	상태성 보통명사	75	'외' 축약 동사
35	시간성 보통명사	76	'외' 축약 형용사
36	수사	77	
37	지시대명사	78	
38	인칭대명사	79	
39	고유명사	80	
40	인칭 고유명사	81	
41	성씨 고유명사	82	

사사전의 각 표제어 정보의 종류와 구성에 대한 다양한 연구가 진행되고 있다. 본 연구에서 개발된 형태소 분석 시스템의 사전 표제어 정보 구성은 다음의 〈표 1〉과 같다.

〈표 1〉에 나타난 바와 같이 본 시스템의 사전 표제어 정보 구성의 특징은 용언 분석 정보와 더불어 각 용언의 불규칙 형태가 표제어의 일부분으로 포함된다는 것이다. 즉, 기존의 형태소 분석 시스템에서 불규칙 용언의 활용꼴을 원형으로 복원함에 있어서 구현 코드 상에서의 규칙을 기반으로 원형 복원이 이루어지는 것과는 달리, 본 시스템에서는 불규칙 활용꼴을 사전의 표제어로 포함시켜 한번의 사전 탐색으로 현재 어절이 어떤 종류의 불규칙 활용꼴인지 여부를 파악할 수 있도록 하였다. 예를 들면, "다르다"는 "르"불규칙으로서 불규칙 활용시에 "달라—"로 변형되므로 이를 사전에 표제어로 추가함으로써 "달라서" 등이 "르" 불규칙의 활용꼴이라는 것을 바로 알 수가 있다.

〈표 2〉의 조사, 어미 사전 정보는 표제어인 조사/어미의 기능에 따라 세부 정보로 표시하고 이를 기반으로 어절 분석 시에 규칙에 의한 제약을 가함으로써 불필요한 분석 결과의 생성을 억제시킨다.

〈표 3〉은 본 시스템에서 적용하고 있는 보조적 연결어미와 보조용언을 나열하고 있다. 보조적 연결어미와 보조용언간의 무의미한 결합은 분석시 많은 과분석 오류를 범하게 된다. 따라서 보조적 연결어미와 보조용언의 결합 여부를 분석하여 이를 결합형 사전으로 구성하면 의미가 연결되지 않는 두 품사간의 결합을 제한시킬 수 있다. 본 연구에서는 다양한 한국어 분석 자료를 바탕으로 보조적 연결어미와 보조용언의 결합 관계를 분석하고 이를 아래의 〈표

〈표 2〉 조사/어미 사전 정보

정 보	내 용	정 보	내 용
1	격조사	1	종속적 연결어미
2	부사격 조사	2	대등적 연결어미
3	관형격 조사	3	어말 어미
4	호격 조사	4	부사격 어미
5	접속격 조사	5	관형격 어미
6	보조격 조사	6	명사전성 어미

〈표 3〉 보조적 연결어미와 보조용언

보조적 연결어미	ㄴ다든지 ㄹ지 게 게끔 고 고는 고만 고자 곤 나 다 다든지 려 려고 려는 아 아다 아야 어 어다 어도 어만 어야 어지지 지 지는 지도 지를 지만도 질
보조 용언	가 계시 겠 나 나서 나오 내 놓 달 달라 대 되 두 드리 들이 듯하 마지않 만 들 만하 말 못하 받 버리 보 보이 싶 싶 않 오 있 좋 주 줄 지 하 한하

4)와 같이 보조적 연결어미-보조용언 결합형 사전으로 구성하였다.

본 연구에서 개발된 시스템에서 사용하고 있는 본용언과 보조용언의 결합형 표제어의 일부분은 〈표 4〉와 같다. 일반적으로 자주 사용되는 보조적 연결어미 30개와 보조용언 36개를 수작업으로 분석하여 의미적으로 결합이 허용되는 결합형 보조적 연결어미-보조용언 238개를 생성하였다. 위의 〈표 4〉에서 표제어의 우측에 표시되어 있는 정보의 형태는 다음의 〈표 5〉와 같다.

보조용언 또한 불규칙 활용이 될 수 있으므로 이를 표시하기 위해서 2자리 숫자를 사용하였다. 첫째 숫자는 불규칙의 종류를 나타내며, 두 번째 숫자는 보조적 연결어미와 보조용언간의 경계 위치를 나타낸다. 이와 같이, 보조적 연결어미와 보조용언의 결합형을 사전으로 구축함으로써 일반 어절 분석에서 도출될 수 있는 다양한 형태의 과분석 오류를 방지할 수 있었다. 예를 들면 보조적 연결어미 "ㄹ지"는 의미적으로 보조용언 "못하"와는 결합할 수 없다. 그러나 보조적 연결어미 "지"는 보조용언 "못하"와 결합이 가능하다. 이러한 의미적 결합관계를 사전에 기술함으로써 "볼지못하다"("보"+"ㄹ지"+"못하"+"다")와 같은 어절이 분석 성공되는 결과를 미연에 방지할 수 있다. 이와 같이 본용언과 보조용언의 분리를 위해 많은 사전 탐색과 규칙 비교 연산 등을 수행함으로써 발생하는 시스템 부하를 줄이고 동시에 분석 정확도를 향상시킬 수 있는 것이다.

나. 사전 접근 API 구현

본 시스템에서 개발된 사전의 구조는 한국어 사전 표제어의 전자적 저장 방법으로 가장 널

〈표 4〉 결합형 보조용언 예

아계시	1	도좋	1
아나	1	려	5
아나가	1	려고	5
아나오	1	려고하	2
아나와	21	려고해	62
아내	1	려는	5
아놓	1	려하	1
아는듯하	2	려해	61
아는듯해	62	아	5
아다	31	아가	1
아다	5	아가지	1
아다드려	12	아가져	11
도	5		

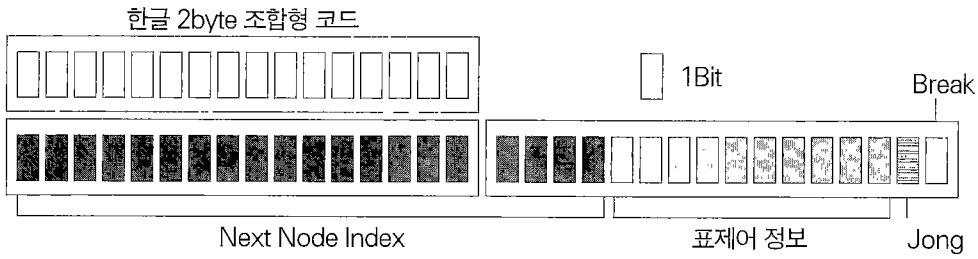
〈표 5〉 결합형 보조용언 사전 정보

정보	내 용
1-4	보조적 연결어미와 보조용언의 경계 위치점
5	보조적 연결어미로만 구성된 표제어
10	"이" 축약
20	"와" 축약
30	"ㄹ" 축약
40	"왜" 축약
50	"워" 축약
60	"해" 축약
70	"ㄷ" 불규칙
비고	-"5"를 제외한 한자리 숫자는 보조용언이 규칙용언인 결합형 용언의 경계점을 표시 -두자리 숫자는 각각 불규칙 종류, 경계점 위치를 표시

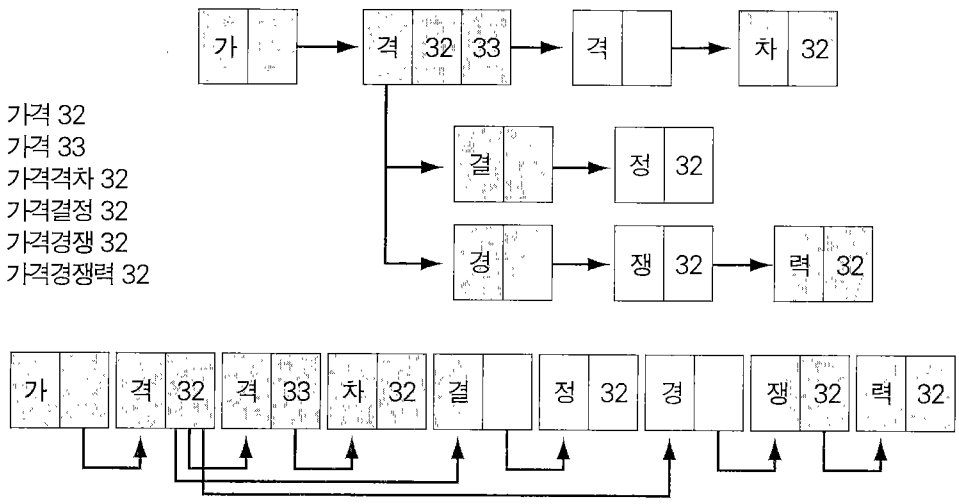
리 활용되고 있는 TRIE에 기반한다. 다음 〈그림 2〉는 TRIE 사전의 노드 구성을 나타낸다.

TRIE 사전의 한 노드는 총 6 byte로 구성된다. 처음 2 byte는 표제어의 한 음절을 저장하기 위해서 사용하고, 현재 음절 노드의 다음 노드를 가리키는 인덱스를 저장하기 위해서 20 bit를 사용한다. 현재 음절 노드가 사전에 등재된 표제어의 마지막 음절일 경우 현재 표제어의 사전 정보를 저장하기 위해서 10 bit를 사용한다. "Jong" 필드는 현재 음절의 종성이 표제어로서 의미가 있는지 없는지를 표시하는 필드로서 정적사전(품사사전)에서는 사용되지 않고, 사용자 정의 사전에서 사용될 수 있다. "Break" 필드는 현재 음절 노드가 표제어의 마지막 음절인지 여부를 표시하는 필드로서 이 필드가 1이면 표제어 사전 정보 필드에 정보값이 저장되게 된다. 〈그림 3〉은 사전 표제어가 실제로 전자사전에 저장되는 구조를 나타낸다.

〈그림 3〉의 상단 부분은 표제어가 전자사전에 저장되는 의미적인 구조를 나타내고 아래 부분은 실제로 메모리와 디스크에 적재되는 구조를 나타낸다. 한 표제어가 여러 개의 사전 정



〈그림 2〉 TRIE 사전 음절 노드의 구조



〈그림 3〉 TRIE 사전에 표제어가 저장되는 구조

```
표제어 입력
표제어 Header Index 생성(첫음절의 완성형 코드값)
if (현재 Header에 삽입된 표제어가 없다.) {
        Header Index 위치에 현재 표제어 음절 노드 생성
} else {
        while (모든 표제어 음절이 다 삽입될 때까지) {
                if (동일 TRIE 레벨에 현재 표제어 음절이 존재한다.)
                        if (찾은 음절 노드가 leaf 노드이다)
                                나머지 표제어 음절을 leaf 노드 다음에 모두 저장
                        else
                                다음 노드로 점프
                else
                                나머지 표제어 음절을 현재 TRIE 레벨에 삽입
        }
}
```

〈그림 4〉 단일 표제어 삽입 알고리즘

```
한글 어절 입력
어절 Header Index 생성(첫 음절의 완성형 코드값)
if (현재 Header에 삽입된 표제어가 없다) {
        입력어절의 첫 음절에서 종성을 제거하고 다시 검색
        if (종성 제거한 음절이 Header에 존재한다)
                이 음절에 대한 모든 사전 정보를 저장하고 리턴
        else
                검색 실패
} else {
        while (모든 입력 어절의 음절이 다 탐색될 때까지) {
                if (동일 TRIE 레벨에 현재 음절이 존재한다) {
                        이 음절까지의 사전 정보를 저장
                        현재 음절의 종성을 제거하고 현재 TRIE 레벨에서 다시 탐색
                        if (존재한다)
                                이 음절까지의 사전 정보를 저장
                        다음 노드로 점프
                } else {
                        현재 음절의 종성을 제거하고 현재 TRIE 레벨에서 다시 검색
                        if (존재한다)
                                이 음절까지의 사전 정보를 저장
                        리턴
                }
        }
}
```

〈그림 5〉 생성된 사전에 대한 탐색 알고리즘

보를 포함하는 표제어의 경우는 각 표제어 정보를 나타내는 음절 노드를 연속해서 저장한다. 이 때 다수의 표제어 정보를 저장하기 위해서 6 바이트 음절 노드 전체를 사용하는데 소요되는 저장 공간 차원에서 비효율성이 문제가 되기는 하나, 사전 크기가 커지더라도 검색 속도를 최적으로 하기 위해 단일화된 노드 접근 차원에서 그 의미가 있다.

TRIE 기반 사전을 생성하기 위한 단일 표제어 삽입 알고리즘은 다음의 〈그림 4〉와 같다.

〈그림 5〉는 생성된 사전에 대한 탐색 알고리즘이다.

사전 접근 API의 근간이 되는 2가지 알고리즘은 실제로 최적의 속도를 위한 단일화된 구

〈표 6〉 단어생성규칙에 따른 어절 분석 기능

	세부 분석 항목	분석 패턴
체언 분석	조사사전 탐색 및 제약조건 검사	
	접미사 분석	
	용언화 접사 처리	"당하-", "시키-", "스러-", "하-", "되-", "롭-", "답-", "있-", "없-", "같-", "치-", "키-"
	"이다" 처리	
용언 분석	복합명사 처리	
	"아/어" 생략 변이체 처리	
	매개모음 "으" 삽입 처리	
	"ㅏ/ㅐ" 변이체 원형 복원 처리	
	규칙 활용꼴 처리	
	불규칙 활용꼴 처리	"ㅅ"불규칙, "ㄷ"불규칙, "ㅂ"불규칙, "ㅎ"불규칙, "우"불규칙, "러"불규칙, "르"불규칙, "으"탈락, "ㄹ"탈락, "하"불규칙, "와"축약, "이"축약, "외"축약
	어미 분석	보조적 연결어미/보조용언 결합, 명사형 전성어미 처리, 선어말 어미 처리, 어미사전 탐색
부사류분석	보조사 사전 탐색	
미등록어 분석	고빈도 조사검사	
	후미어 패턴 검사	
수사 분석	수사 패턴 검사	
	단위명사 분석	
	분석 모드에 따른 후미어 형태소 분석	

조를 가진다. 특히 사전 탐색 알고리즘에서 현재 입력 어절의 사전 탐색 결과로 모든 하위 스트링에 대한 사전 탐색 결과를 함께 도출해야 하므로 효율적인 구현이 절대적으로 요구된다. 또한 명사류 정보에 대한 일괄적인 처리를 위해서 사전 탐색 알고리즘에서는 모든 종류의 명사 정보를 하나로 묶어서 명사 표시 정보를 리턴하는 기능이 있다.

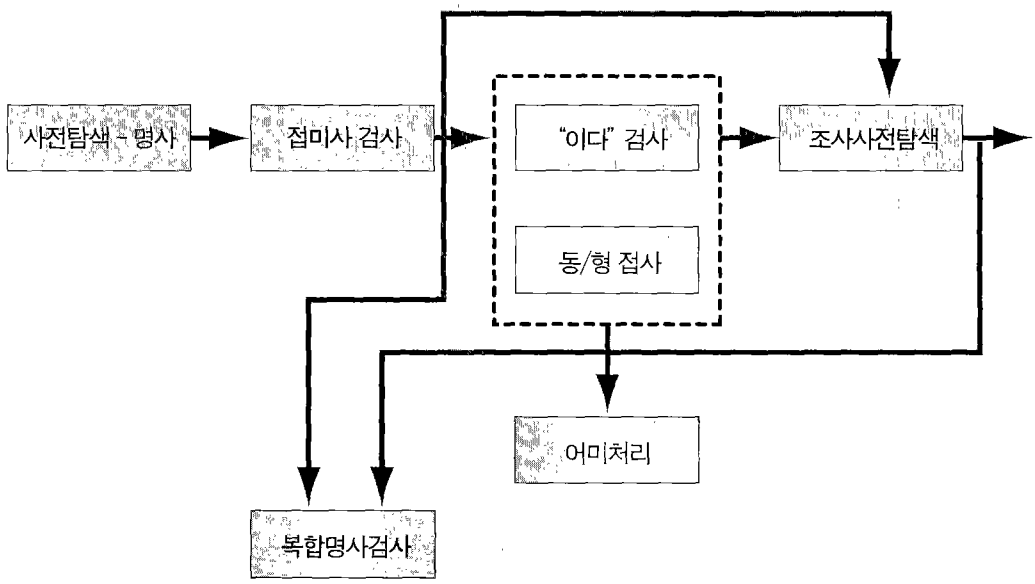
4.2 어절 형태소 분석

입력 어절에 대한 형태소 분석은 크게 (1) 체언 분석, (2) 용언 분석, (3) 부사류 분석, (4) 독립언 분석, (5) 미등록어 분석 등으로 나뉜다. 체언 분석은 조사 사전 탐색 및 제약 조건 검사, 접미사 처리, 용언화 접사 처리, "이다" 처리, 복합명사 처리 등으로 구분된다. 체언 분석 중에서 용언화 접사 처리 부분은 체언 분석에서 용언 분석 부분인 어미 처리로 분석 특성이 변화되는 특징이 있다. 용언 분석에는 어미 처리, 선어말 어미 처리, 보조용언 처리 등이 있으며, 다음 〈표 6〉는 단어형성규칙에 따른 세부 어절 분석 기능을 열거하고 있다.

미등록어 분석은 자주 사용되고 적용 패턴이 일정한 조사가 미등록어에 존재하는 가에 대한 여부를 검사하는 부분과 대용량 말뭉치에서 추출된 체언의 후미어 패턴 사전에 기반한 후미어 패턴 검사로 나뉜다. 수사 분석은 수사 형태소 패턴에 따른 수사, 수관형사, 단위명사, 후치 명사 등을 판별하는 기능으로 구성된다.

가. 체언 분석

문서의 특징을 나타내는 색인어의 대부분을 이루는 명사, 대명사, 고유명사 등에 대한 어절 분석은 자동 색인 시스템의 중요한 부분이다. 특히 여러 개의 명사가 결합되어 있는 복합 명사 분석은 전체 시스템 속도의 약 50% 이상을 차지하므로 분석 속도를 최적화하기 위한 다양



〈그림 6〉 체언 분석 과정

한 방법론이 적용되어야 한다. 아래의 〈그림 6〉은 사전에서 체언으로 판명된 좌측 최장 부분 어절의 나머지 부분 어절을 분석하는 과정을 도식화한 것이다.

사전 탐색에서 명사로 판별된 어절에 대해서 접미사 검사를 한 후, "이다" 조사가 붙거나 동사/형용사화 접사가 붙은 어절에 대해서 어미 처리를 수행한다. 또한 각 단계별로 제약조건을 두어서 각 분석 단계에 들어갈 필요가 없는 부분 어절에 대해서는 바로 다음 단계로 넘어갈 수 있도록 한다. 조사사전 탐색 부분은 조사 바로 앞 음절 정보와 조사 부분 어절 정보를 입력받아서 체언과 조사의 결합 제약조건을 검사한다. 또한 체언 분석 부분은 접미사 검사나 다른 부수적인 검사 루틴에서 모두 실패한 부분 어절에 대하여 복합명사 분석을 수행한다. 본 시스템에서는 다음과 같은 두 가지의 가정 하에서 복합명사 분석을 수행하였다.

가정 1. 접미사가 붙은 명사는 복합명사의 부분 명사가 될 수 없다.
- 복합명사를 구성하는 명사들 중 가장 마지막에 위치한 명사에는 접미사가 붙을 수 있다.

```
S_sub : 입력 어절에서의 현재 분석 대상 부분 어절
U_i : 분석 대상 음절
S_sub = U_i U_i+1 U_i+2 ... U_i+k

for (j = 0 to k-1) {
        if (j == 1)
                continue;
        U_i+j 위치에서 사전 탐색 수행;
        if (2음절 명사) {
                VERTEX[Index]에 현재 사전 탐색 결과 저장;
        }
}

연속된 명사 리스트를 찾아서 DADJ에 인덱스를 저장;
while (VERTEX를 다 검사할 때까지) {
        PUSHSTACK <-- Initial Index;
        while (!StackEmpty) {
                POPSTACK;
                현재 명사 노드 정보를 형태소 분석 결과 버퍼에 저장;
                현재 노드와 연관되는 명사 리스트 정보를 DADJ에서 추출하여 스택에 저장;
                접미사 검사;
                "이다" 조사 검사;
                용언화 접사 검사;
                조사 검사;
        }
}
```

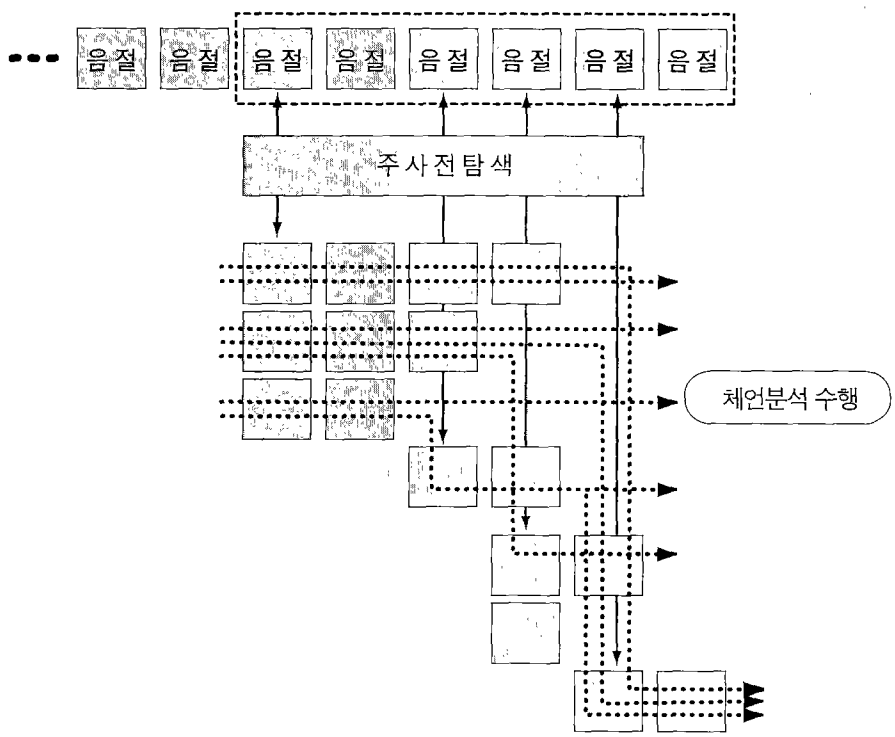
〈그림 7〉 복합명사 분석 알고리즘

가정 2. 한 글자로 구성된 명사는 복합명사의 부분 명사가 될 수 없다.

이와 같은 가정 하에서 입력 어절에 대한 복합명사의 분석은 다음의 〈그림 7〉과 같은 알고리즘으로 수행된다.

우선, 남은 분석 어절의 모든 위치에서 사전 탐색을 수행하여 2음절 이상의 명사를 추출하고, 추출된 2음절 이상의 명사에 대한 현재 분석 어절에서의 위치와 함께 사전 정보를 하나의 vertex로 구성한다. 이렇게 구성된 vertex에 각각의 명사 위치와 길이를 검사하여 연결된 명사 리스트를 생성하고, 이를 매트릭스에 저장한다. 실제 복합명사 분석 모듈에서는 스택을 이용하여 유효한 명사 리스트를 따라가며 분석을 수행한 후, 복합 명사 분석 결과를 생성하게 된다. 다음의 〈그림 8〉은 위의 알고리즘을 이용하여 복합명사가 분석되는 개념과 구조를 나타낸다.

본 연구에서 수행된 복합 명사 분석 알고리즘은 모든 분석 대상 음절 위치에서의 사전 탐색에 따른 사전 탐색 횟수의 증가에 대한 문제점이 있을 수 있으나, 일반적인 복합명사 분석 알고리즘에서 사용하고 있는 재귀적 호출이나 복잡한 모듈의 구현을 피할 수 있는 장점이 있다. 그리고 14만 어절/sec의 사전 탐색 속도를 가지는 사전 탐색 알고리즘으로 속도에 대한 문제점을 극복하였으며, 오히려 기존 복합명사 분석 알고리즘보다 빠른 분석 속도를 유지하도록 하였다.



〈그림 8〉 복합명사 분석 과정

나. 용언 분석

자동 색인 시스템에서의 용언에 대한 분석은 두 가지 측면에서 중요한 의미를 가진다. 첫째, 용언 분석 기능이 미약하면 분석 과정에서 처리되지 못한 용언들이 미등록어 처리기로 넘어가게 되고 그 결과 무의미한 색인어의 과도한 생성이라는 문제점이 발생한다. 둘째, 기존의 정보검색 시스템이 체언 중에서도 명사만을 색인어로 채택한 것과는 달리, 본 연구에서는 동사나 형용사까지도 색인어로서 추출하여 문장의 특성을 나타내는 색인어 리스트를 생성할 수 있도록 하였다. 그 이유는 정확한 용언 분석은 시스템 전체의 확장성과 밀접한 관계가 있기 때문이다.

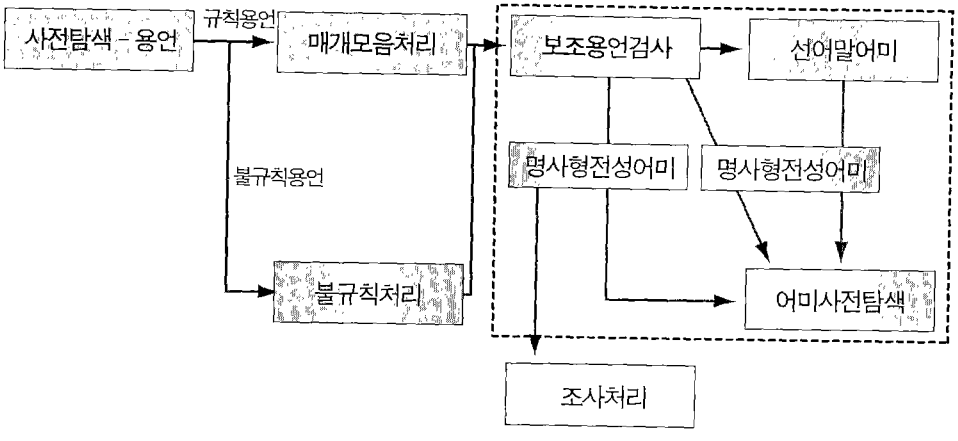
다음의 〈그림 9〉는 문장 내에 출현하는 용언에 대한 분석 과정을 나타내고 있다. 사전 탐색에서 좌측 최장 부분 어절이 용언으로 판단되면 우선 선택된 용언이 규칙인지 불규칙인지를 결정한다. 이러한 결정은 상기에서 언급된 불규칙 사전 정보에 의하여 즉시 이루어진다. 불규칙 용언으로 판정이 되면 불규칙 처리 루틴을 통하여 원형을 복원하고 어미처리부의 시작 부분인 보조용언 검사를 수행하게 되며, 검사 후에 선어말 어미 검사를 수행하고 어미 사전 탐색을 하게 된다. 이때 보조용언 검사 후나 선어말 어미 검사 후에 명사형 전성어미('ㅁ", '음", "기")가 감지되면 체언 분석 루트로 넘어가서 조사 사전 탐색을 수행하게 된다.

다. 부사, 독립어, 관형사 분석

본 연구를 통하여 개발된 분석 모듈은 부사 다음에 올 수 있는 형태소는 보조사 외에는 없다고 가정하고 보조사 사전을 탐색한다. 독립어는 홀로 존재하게 되므로 뒤에 다른 어떤 형태소도 올 수 없다. 관형사는 뒤에 명사 및 복합 명사가 올 수 있으므로, 나머지 분석 어절에 대해서 복합명사 분석을 수행하게 된다.

라. 수사 형태소 분석

본 연구에서 개발된 수사 형태소 분석 모듈은 모든 수사 패턴들을 처리하기 위하여 일반



〈그림 9〉 용언 분석 과정

적인 분석 모듈이라기 보다는 일반 문서에서 높은 빈도로 출현되는 수사 패턴에 대한 효율적인 처리를 위한 모듈이다. 따라서 본 모듈은 다어절에 걸친 복합 수사를 처리하기 위해서 전처리 단계로 태깅 모듈을 수행하거나 구문 분석을 통한 복합적인 처리보다는 한 어절 내에 포함되어 있는 수를 의미하는 단어를 인식하고 이를 효과적으로 분석하거나 색인어로 추출하는 기능을 수행한다(채영숙 외 1997). 〈그림 10〉은 수사 형태소의 분석을 위한 예이다.

2000년
칠천팔백오십육만원을
수백만명이
일학년
1천5백3십5가지

〈그림 10〉 수사 패턴 예

일반적인 표제어 사전과는 달리 수사 사전은 기능적으로 제한되어 있으며, 형태가 일정하게 유지된다는 특징이 있다. 따라서 본 연구에서는 수사 형태소 분석을 위한 특화된 사전 정보와 사전 표제어를 구성하고 이를 시스템에 적용하였다.

〈표 7〉은 수사의 기능과 형태에 따른 수사 표제어 정보를 나타낸다. 단위명사 결합형 수사는 사용빈도가 아주 높은 수사와 단위명사의 결합형을 사전에 추가함으로써 분석시 발생하는 문제점들을 해결할 수 있도록 수사 사전에 추가하였다. 예를 들면, "넉달", "닷냥" 등은 수사가 뒷부분의 단위명사와 결합하면서 변형이 생긴 어절이다. 따라서 이를 구분하여 수사 사전과 단위명사 사전에 추가하기보다는 하나로 묶어서 루틴에서 처리를 하는 것이 더 효율적이다. 왜냐하면 "넉", "닷"과 같은 변형 수사를 수사 사전에 모두 추가하면 이를 위한 제약 조건 검사도 훨씬 많아지기 때문이다. 이 사전정보를 바탕으로 약 91개의 수사 표제어와 263개의 단위명사를 정제하여 사전으로 구성하였으며, 수사 형태소 분석을 수행하기 위하여 〈그림 11〉과 같은 분석 알고리즘이 적용되었다.

수사 뒤에 나오는 조사나 접미사 혹은 명사의 분석은 일반 어절 형태소 분석 모듈을 수정하여 분석 모드에 따른 어절 분석이 가능하도록 하였다. 즉, 수사 다음의 조사나 접미사를 분석하기 위해서는(SUSA_POSTNOUN 모드) 일반 명사가 사전에서 검색되었다고 가정

〈표 7〉 수사 사전 정보

101	:	한자표기 수사 (일, 이, 삼,...,십)
102	:	한글표기 수사 (하나, 둘, 셋,..., 열, 스물, ..., 아흔)
103	:	불완전형 한글표기 수사 (한, 두, 세, 네)
104	:	단위명사 결합형 (자주 쓰이는 수사 + 단위명사)
105	:	바로 뒤에 항상 단위전치어/단위명사가 오는 수사
106	:	"수"가 앞에 붙어서 불특정 수/양을 나타내는 수사
107	:	백, 천, 만, 억 (101, 102와 같이 결합하여 사용)

```
f (어절이 숫자로 시작) {
        앞부분의 모든 숫자를 수사 형태소 분석 버퍼로 저장;
}
if (어절 길이가 1 이하) {
        분석 실패;
}
수사 사전 탐색;
if (어절이 수사로 시작) {
        수사를 수사 형태소 버퍼에 저장;
}
while (1) {
        if (어절의 마지막)
                break;
        단위명사 분석;
        조사/접미사 분석;
        수사 다음의 명사 분석;
        현재 분석 위치에서 수사 사전 탐색;
        if (탐색 성공) {
                수사 사전 정보에 따른 규칙 적용;
                수사 형태소 분석 버퍼에 저장;
        } else {
                단위명사 분석;
                조사/접미사 분석;
                수사 다음의 명사 분석;
        }
        분석 위치 이동;
}
if (어절이 수관형사로 시작) {
        수관형사 뒤에는 십, 백, 천, 만, 억, 단위명사만이 온다고 가정;
        "여러" 다음에는 무조건 단위명사만 온다.;
        단위명사 분석;
        리턴;
}
```

〈그림 11〉 수사 형태소 분석 알고리즘

하고 명사 다음의 조사나 접미사를 분석하듯이 수사의 후절어를 분석해야 하며, 수사 다음의 명사 분석(SUSA_NOUN 모드)도 동일하다. 이와 같이 분석 모드에 따른 다양한 어절 분석 API를 제공함으로써 보다 세부적인 수사 형태소 분석을 수행할 수 있다.

5 실험 및 결과 고찰

대부분의 정보검색 시스템이 대용량의 데이터를 효과적으로 처리하기 위한 고성능 서버에서 수행되므로 이 시스템의 하부 시스템으로 탑재되는 자동 색인 시스템도 대용량, 고성능 서

버에서의 속도나 정확도가 더 의미가 있다. 따라서 본 연구에서 개발된 어절 분석 시스템의 성능 실험 환경을 〈표 8〉과 같이 구성하였다.

또한 어절 분석 즉, 형태소 분석 결과의 정확도를 측정하기보다는 어절 분석 시스템의 윗부분에 색인어 필터링 시스템을 탑재한 한국어 어절 색인 시스템의 결과를 바탕으로 실험을 하였다. 완벽한 색인 시스템을 위해서는 품사 태거나 구문 분석 태거가 구성되어야 하겠으나 미등록어 처리 및 기타 속도 문제 때문에 실제적으로 적용된 예가 드물다. 따라서 이 논문에서는 어절 분석 결과에서 적절한 색인어를 필터링하는 시스템을 구현하였다.

5.1 분석 속도 측정

분석 속도 측정은 두 가지 시스템을 통하여 수행되었다. 시스템 B는 고빈도 말뭉치에서 추출한 기분석 사전(80만 어절)을 적용한 시스템이고 시스템 A는 적용하지 않은 시스템이다. 다시 말해서 시스템 B는 이 논문에서 구현된 시스템 A에 기분석 사전과 그 탐색 알고리즘이 적용된 시스템이다. 이 시스템에 적용된 기분석 사전은 원시 말뭉치 6,000만 어절에서 추출한 상위 빈도 순위 80만 어절을 시스템 A에 적용시켜서 구성한 사전이다. 시스템 B를 속도 측정 대상에 포함시킨 이유는 기분석 사전의 적용이 전체 어절 분석 시간에 미치는 영향이 어느 정도인가를 알아보기 위해서이다. 분석 대상 데이터는 2000년도 초등학교 5학년 교과서(513,922어절) 데이터이다.

〈표 9〉에서 시스템 B는 시스템 A보다 실제적인 어절 분석 속도가 40% 높음에도 불구하고 전체적인 분석 시간은 크게 차이가 없음을

〈표 8〉 어절 분석기반 자동색인 시스템의 성능 실험 환경

분석환경	O/S : 아델 리눅스 CPU : Pentium Ⅲ 933*2 M/M : 1Gbyte
대상자료	2000년도 초등학교 5학년 교과서(513,922어절) 한국경제신문 데이터(1,000,000어절)

〈표 9〉 어절 분석 속도

	시스템 A	시스템 B
분석 사전 메모리 적재 시간	0.3(초)	3(초)
어절 분석 시간	8.71(초)	5.29(초)
총 분석 시간	9.01(초)	8.29(초)
단위 초당 분석 어절 수	**57,039.07(어절/초)**	**61,933(어절/초)**
말뭉치 Hit Ratio	-	31.4%

알 수 있다. 이는 시스템 B가 기분석 사전을 메모리로 적재하는데 드는 시간이 있기 때문이다. 현재 실험환경에서의 분석 사전 및 기분석 사전을 메모리로 적재하는데 드는 시간이 3초이므로 이 시간은 분석 어절 개수와는 상관없이 일정하다. 그러므로 현재보다 더 많은 양의 데이터를 분석할 때는 기분석 사전을 탑재한 자동 색인 시스템이 유리하다.

5.2 분석 정확도 측정

초등학교 5학년 교과서와 한국경제신문 데이터에서 임의로 각각 1,000어절을 추출하여 색인어 추출 성공률을 계산하였다. 색인어 추출 성공률은 다음 식에 의해서 계산된다.

$$\text{색인어 추출 성공률} = \frac{\text{올바로 추출된 색인어 개수}}{\text{추출된 전체 색인어 개수}}$$

다음의 〈표 10〉은 데이터의 종류에 따른 분석 정확도를 나타낸다.

분석 정확도 측정은 국문학 전공 평가자 3명이 같은 결과 데이터를 각각 평가함으로써 이루어졌다. 같은 어절 분석 결과에 대한 서로 다른 평가 결과는 평가자 3명과 평가 관리자 1명이 그 결과에 대한 의견 조율을 함으로써 재평가 되었다. 교과서 데이터보다 한국경제신문 데이터에서 정확도가 다소 떨어짐을 알 수 있다. 이러한 결과는 한국경제신문 데이터에서 전문용어나 미등록어가 다수 발생하여 사전 표제어 부재로 인한 분석 실패에서 그 원인을 찾을 수 있다. 분석 오류나 색인어 추출 실패의 요인은 세 가지로 나뉜다. 첫째, 고유명사와 긴 후미어가 결합되어 있는 경우에는 후미어가 완벽하게 분리되지 않는 경우가 많았다. 따라서 신조어나 고유명사에 대한 사전 표제어 보강과 함께 체언 후미어 처리에 대한 다양한 규칙 학습이 필요하다. 둘째, "-부터가"와 같은 복합조사나 "-이야요", "-이예요"와 같은 구어체 어미에 대한 처리가 미흡하다. 셋째, 복합명사 분석에 있어서 불필요한 명사 분리가 발생한다. 예를 들면, "구체적인"—〉"구체" + "적인", "오염물질의"—〉"오염물" + "질의"와 같은 복합명사 분리는 규칙을 적용함으로써 방지가 되었지만 "관자놀이"가 사전에 존재함에도 불구하고 "관자놀이의"—〉"관자" + "놀이" + "의"와 같은 복합명사 분리는 의미가 없다. 따라서 사전에 존재하는 복합명사를 특수하게 처리하여 우선적으로 분석된 결과가 사전에 표제어로 등록되어 있는 복합명사이면 그 복합명사를 분리하여 분석 결과를 도출하는 것을 막음으로써 문제를 해결하였다.

6 결론 및 향후 연구

본 연구에서는 정보검색시스템의 성능향상을 위하여 기존에 연구되었던 다양한 어절 분

〈표 10〉 분석 정확도

분석환경	교과서	한국경제신문	평균
정확도	98.7842 (%)	87.9013 (%)	93.3427 (%)

석 기법들을 바탕으로 어절분석 속도의 최대화, 형태소 분석기의 모듈화 및 구조화 그리고 형태소의 정확한 분석을 위한 한국어 어절 분석 시스템을 개발하였다.

본 연구에서 개발된 시스템의 특징은 어절 분석 속도를 높이기 위하여 품사 사전의 구조화와 탐색 방법에 대한 다양한 접근 방법의 평가를 통해 최적의 알고리즘을 구현하였으며, 전체적인 시스템 구조를 디자인함에 있어서 완벽하게 모듈화된 하부 시스템의 유기적이고 효율적인 결합에 중점을 두고 각 모듈별 성능 및 속도 검증이 가능하도록 하였다. 또한, 대부분의 형태소 분석 시스템에서 적용하고 있는 재귀적 복합명사 분석을 탈피하여 빈번한 재귀적 호출에 따른 시스템 부하를 줄이고 확장성을 도모하였으며, 다층적 수사 패턴 인식에 기반한 수사 형태소 분석 시스템을 개발하여 형태소 분석 시스템과 결합하였다. 마지막으로 다어절 분석을 위한 기반 구조를 제공하기 위해서 원형 큐 기반 다어절 토큰 관리기를 개발하여 확장성을 극대화하였으며, 고성능 서버환경에서 대용량 문서 집합을 색인할 때, 어절 분석 속도를 향상시키기 위하여 해싱 기반 기분석 사전 탐색 및 필터링 모듈이 구현되었다.

본 연구에서 구현된 형태소 분석 시스템이 정보검색 시스템과 결합하여 검색 결과의 정도율과 재현율을 최적화시키기 위해서는 부가적인 시스템이 추가로 결합되어야 한다. 우선 검색의 대상이 되는 많은 문서가 띄어쓰기, 철자 등을 포함한 다양한 어절 기반 오류를 포함하고 있기 때문에 오류가 포함된 문서를 효과적으로 색인하기 위해서는 자동 띄어쓰기 기능과 철자 교정 기능 등이 자동 색인 엔진과 유기적으로 결합해야 한다. 이러한 문제를 해결하기 위하여 다양한 접근 방법이 시도되었으나 기존의 형태소 분석 시스템은 색인 시스템과 기타 어절 오류 분석 시스템을 독립적으로 결합시키기 때문에 두 종류의 시스템이 서로 교환하고 공유해야 하는 많은 어절 정보들이 소실되게 된다. 따라서 형태소 분석기와 어절 오류 분석기를 유기적으로 통합하는 작업이 요구된다. 이를 위해서 음절 N-gram기반 자동 띄어쓰기 오류 수정 시스템과 오류 패턴에 기반한 철자 오류 수정 시스템을 개별적으로 개발하여 성능을 검증하였으나, 이를 유기적으로 통합하는 작업은 향후의 연구로 남겨 두었다.

참고문헌

강승식. 1993. 『음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석』. 서울대학교 컴퓨터공학과 박사학위 논문.

강승식. 1993. 음절 특성을 이용한 한국어 불규칙 활용 어절의 형태소 분석 방법, 『1993년도 제5회 한글 및 한국어 정보처리 학술발표 논문집』.

강승식. 1994. 다층 형태론과 한국어 형태소 분석 모델. 『제6회 한글 및 한국어 정보처리 학술발표 논문집』. pp.140-145.

김민정. 1997. 『규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거』. 부산대학교 전자계산학과 박사학위 논문.

『동아 새국어사전』. 1995. 서울: 동아 출판사.

신기철, 신용철. 1994. 『새우리말 큰사전』. 삼성출판사.

이영식. 1994. 『사전 근사탐색과 Heuristics를 이용한 한국어 철자 오류 교정 시스템 구현』. 부산대학교 전자계산학과 석사학위 논문.

채영숙, 김재원, 김민정, 권혁철. 1991. 한국어 철자 검색을 위한 형태소 분석 기법. 『91 우리말 정보화 잔치』. 국어정보학회, 179-186.

채영숙. 1998. 『연어 규칙에 기반한 한국어 문서 교정시스템의 구현』. 부산대학교 전자계산학과 박사학위 논문

Baeza-Yates, Ricardo, and Ribeiro-Neto, Berthier. 1999. *Mordern Information Retrieval*. New York: ACM Press.

Cahill, L.J. 1990. "Syllable-based Morphology." *Proceedings of the 13th International Conference on Computational Linguistics*. Vol.3, 48-53.

Charniak, Eugene. 1993. "Statistical Langual Learning." *A Bradford Book*. Cambridge: The MIT Press.

Kang, S.S. 1992. "A Statistical Approach to Syllable-based Morphological Analysis", *Proceedings of the International Conference on Computer Processing of Chinese and Oriental Language*.

Kelly, Douglas G. 1994. *Introduction to Probability*. London: Macmillan Publishing Company

Koskcnniemi, K. 1983. "Two-level Model for Morphological Analysis," *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, 683-685.

Kwon, H.C., Chae, Y.S. and Jeong, G.O. 1991. "A Dictionary-based Morphological Analysis," *Proceedings of NLPRS '91*, 87-91.

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

Manning, Christopher D. and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.