자동 색인을 위한 한국어 형태소 분석기의 실제적인 구현 및 적용

최 성 필[†] · 서 정 현[†] · 채 영 숙[††]

요 약

본 논문에서는 정보검색 시스템에서 필수적인 자동 색인을 위한 한글 형태소 분석기를 구현하였다. 현존하는 대용량의 데이터에 대한 자동 색인을 효율적으로 수행하기 위해서 새로운 개념이나 아이디어의 도입 및 적용에 초점을 맞추기보다는 기존에 연구되었던 다양한 어절 분석 기법들을 바탕으로 어절분석 속도의 최대화, 형태소 분석기의 모듈화 및 구조화에 초점을 맞추었다. 따라서 본 논문에서 개발된 시스템의 특징은 이론적인 측면보다는 소프트웨어 공학적인 측면이 훨씬 더 강조된다. 품사 사전의 구조화가 우선적으로 수행되었으며, 이에 따라서 체언 및 용언 분석 모듈, 수사 분석 모듈 등이 구현되었다. 또한 형태소의 패턴을 이용한 미등록어 분석 기능이 개발되었다. 개발된 전체 시스템은 정보검색 엔진인 K-2000 시스템의 색인 모듈로 장착되어서 적용되었다.

Practical Development and Application of a Korean Morphological Analyzer for Automatic Indexing

Sung-Pil Choi[†] · Jerry Seo[†] · young-suk Chae[††]

ABSTRACT

In this paper, we developed Korean Morphological Analyzer for an automatic indexing that is essential for Information Retrieval. Since it is important to index large-scaled document set efficiently, we concentrated on maximizing the speed of word analysis, modularization and structuralization of the system without new concepts or ideas. In this respect, our system is characterized in terms of software engineering aspect to be used in real world rather than theoretical issues. First, a dictionary of words was structured. Then modules that analyze substantive words and inflected words were introduced. Furthermore numeral analyzer was developed. And we introduced an unknown word analyzer using the patterns of morpheme. This whole system was integrated into K-2000, an information retrieval system.

1. 서 론

최근 들어 다양한 형태소 분석 시스템이 개발되어 활용되고 있으며, 실제로 많은 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, 시스템 개발자 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 시스템 관리자 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 논문에서는 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 논문의 목적은 새로운 알고리즘이나 아이디어를 개발하는 것이 아니다. 이미 형태소 분석 수준에서는 많은 유능한 연구자들에 의해서 다양한 분석 기법과 아이디어가 많이 개발되어 있다. 문제는 이러한 효과적인 기법들이 어떻

† 정 회 원 : 한국과학기술정보연구원 연구원
†† 정 회 원 : 영산대학교 멀티미디어공학부 교수
논문접수 : 2002년 8월 22일, 심사완료 : 2002년 9월 14일

게 유기적으로 결합되느냐이다. 본 논문의 목적은 이런 기법들의 적용과 결합에 있다. 예를 들어, 어절 분석에는 최장일치에 기반한 좌-우 분석과 규칙 패턴과 사전에 기반한 우-좌 분석이 존재한다[1, 5]. 좌-우 분석은 한국어 어절의 생성 오토마타의 순서를 그대로 적용하므로 구현이나 관리에 편리하다. 우-좌 분석은 형식형태소를 우선적으로 분석함으로써 사전 탐색 회수를 줄일 수 있고 미등록어에 대한 방어 능력을 강화할 수 있다[1, 5]. 따라서 이 두 가지 기법을 하나로 결합하면 두 기법들의 장점을 동시에 얻을 수 있다. 본 논문에서 개발된 시스템은 이러한 유용한 기법들의 결합과 시스템의 튜닝 및 관리를 위한 디자인 및 구현, 그리고 형태소 분석기가 수행해야 하는 필수적인 기능들을 모두 구현하고 이를 결합하는 형태로 개발되었다.

이 논문은 다음과 같이 구성된다. 우선 2장에서 형태소 분석 기술의 특징과 일반적인 언어 처리 기능 외에 필요한 필수 요소를 정의하고, 3장에서는 개발된 시스템의 전체적인 구조와 특징을 설명한다. 4장에서는 각 모듈별로 형태소 분석 기능이 어떻게 구성되고 구현되었는가를 설명하고 5장에서는 본 논문에서 개발된 시스템이 정보검색시스템에 적용되는 방법에 대해서 기술한다. 마지막으로 결론과 향후 연구 방향에 대해서 논의한다.

2. 형태소 분석기술의 특징

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

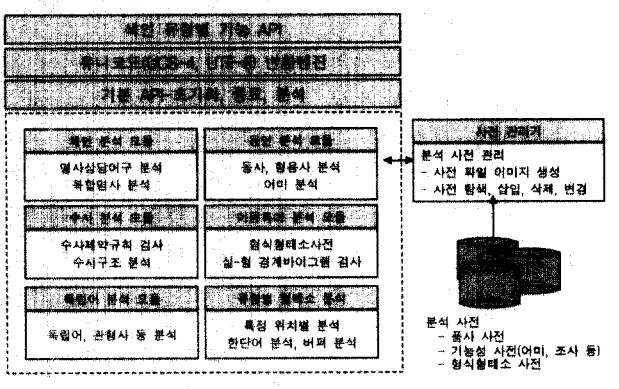
<표 1>은 본 논문에서 제시한 자동 색인을 위한 형태소 분석 시스템의 필수 요구사항을 보여준다. 큰 규모의 언어처리 시스템과 정보검색 시스템의 기반 시스템으로의 역할을 충실히 수행하기 위해서 본 논문에서는 기본적인 언어학적 분석 외에 다음과 같은 요구사항을 정의하고 이러한 다양한 요구사항에 적절히 대처할 수 있는 시스템을 개발하려고 하였다.

〈표 1〉 형태소 분석 시스템의 필수 요구사항

항 목	세 부 요 구 사 항
사 전	어절 분석 속도를 높이기 위하여 품사 사전의 구조화와 탐색 방법, 또한 이를 위한 다양한 접근 방법에 대한 평가를 통해 최적의 알고리즘이 구현되어야 한다.
모 듈 화	전체적인 시스템 구조를 디자인함에 있어서 모듈화된 하부 시스템이 유기적으로 결합되어야 하고 모듈별로 차별화된 검증 및 평가가 가능해야 한다.
복합명사	대부분의 형태소 분석 시스템에서 적용하고 있는 재귀적 복합명사 분석을 탈피하여 빈번한 재귀적 호출에 따른 시스템 부하를 줄이고 확장성을 도모할 수 있어야 한다.
수 사	한국어 어절의 상당부분을 차지하고 있는 수사 분석이 가능해야 한다.
미등록어	형태소 분석에 실패한 어절들에 대한 적절한 미등록어 형태소 추정이 가능해야 한다.

3. 형태소 분석 시스템의 구조

본 논문에서 개발한 형태소 분석 시스템의 전체적인 구조는 (그림 1)과 같다. (그림 1)에서 보는 바와 같이 각 기능적 요소들은 완벽하게 모듈화되어 있다. 또한 각 모듈별로 특수 기능을 수행하는 하부 모듈이 존재한다. 위 그림에서 보이는 각 모듈은 형태소 분석 상에서 가장 크게 그 기능을 차지하는 모듈이다. 만일 시스템 상에서 새로운 기능이 필요하다면 그 기능에 부합하는 모듈을 구현하여 형태소 분석 모듈 패키지의 일부분으로 추가하고 기존의 다른 모듈과 연동하는 과정만을 기술하여 적용시키면 된다. 이 작업이 용이한 이유는 각 모듈간의 통신이 형태소 분석 결과를 저장하는 전역변수 1개와 함수 파라메터, 리턴 값으로만 이루어지기 때문이다. 형태소 분석 모듈 패키지를 둘러싸는 'WRAPPER 기능 모듈'은 사용자의 요구사항이나 시스템 적용 환경에 맞도록 형태소 분석 모듈들을 이용하여 다양한



(그림 1) 형태소 분석 시스템 및 색인 시스템 구조

기능을 구현하는 기능을 수행한다. 사전 관리기는 Tree 구조를 기반으로 시스템에서 사용하는 사전들을 저장하고 이를 고속으로 접근할 수 있는 API를 제공하며, 실제 어절 분석 모듈인 어절 형태소 분석기는 한국어 어절의 구성 형태에 따른 유한 오토마타의 각 상태 전의 구조로 이루어진다. (그림 2)는 형태소에 대한 중간 분석 결과 및 최종 분석 결과를 저장하는 구조체이다.

```
/* 형태소/정보 리스트 구조체 */
typedef struct {
    UWORD Morpheme[MORP_LEN] ;
        /* 형태소, 최대 크기 20자 */
    UWORD ninfo ;
        /* 사전 정보 개수 */
    BYTE info[MORP_INFO] ;
        /* 사전 정보 */
} tMORP_ITEM ;
/* 형태소 분석 결과 저장 구조체 */
typedef struct {
    UWORD nMorp ;
        /* 한 어절에 대한 형태소 수 */
    tMORP_ITEM MI[MAX_MORP_LIST] ;
        /* 형태소 리스트 */
} tMORP_RESULT ;
```

(그림 2) 형태소 분석 결과 저장 구조체

위에서 설명된 하부 모듈들은 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. 한국어 형태소 분석 시스템

4.1 어절 분석 사전의 구조와 사전 접근 API

4.1.1 어절 분석 사전의 구조

본 논문에서 개발된 형태소 분석 시스템의 사전 표제어 정보 구성은 부록의 <표 2-1>와 같다. <표 2-1>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 포함시킴으로써 한번의 사전 탐색으로 현재 어절이 어떤 종류의 불규칙 활용꼴인지 여부를 즉시 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 2> 조사/어미 사전 정보

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

<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 논문에서는 다양한 한국어 분석 자료를 바탕으로 보조적 연결어미와 보조용언의 결합 관계를 분석하고 이를 보조적 연결어미-보조용언 결합형 사전으로 구성하였다.

<표 3> 보조적 연결어미와 보조 용언

보조적 연결어미	ㄴ다든지 ㄹ지 게 개끔 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
보조용언	가 계시 겠 나 나서 나오 내 놓 달 달라 대 되 두 드리 들이 듯하 마지않 만들 만하 말 못하 받 버리 보 보이 싶 싶 않 오 있 좋 주 줄 지 하 한하

일반적으로 자주 사용되고 있는 보조적 연결어미 30개와 보조용언 36개를 수작업으로 분석하여 의미적으로 결합이 허용되는 결합형 보조적 연결어미-보조용언 238개를 생성하였다. 보조적 연결어미와 보조용언 결합형 사전의 사전 정보는 <표 4>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 "불지못하다"("보"+"ㄹ지"+"못하"+"다")와 같은 어절이 분석 성공되는 결과를 미연에 방지할 수 있다. 이와 같이 본용언과 보조용언의 분리를 위해 많은 사전 탐색과

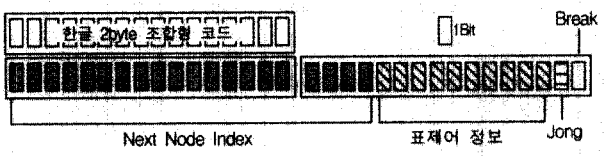
규칙 비교 연산 등을 수행함으로써 발생하는 시스템 부하를 줄이고 동시에 분석 정확도를 향상시킬 수 있는 것이다.

〈표 4〉 결합형 보조용언 사전 정보

정 보	내 용
1-4	보조적연결어미와 보조용언의 경계 위치점
5	보조적 연결어미로만 구성된 표제어
10	"이" 축약
20	"와" 축약
30	"르" 탈락
40	"왜" 축약
50	"워" 축약
60	"해" 축약
70	"ㄷ" 불규칙
비 고	• "5"를 제외한 한자리 숫자는 보조용언의 규칙 용언인 결합형 용언의 경계점을 표시 • 두자리 숫자는 각각 불규칙 종류, 경계점 위치를 표시

4.1.2 사전 접근 API 구현

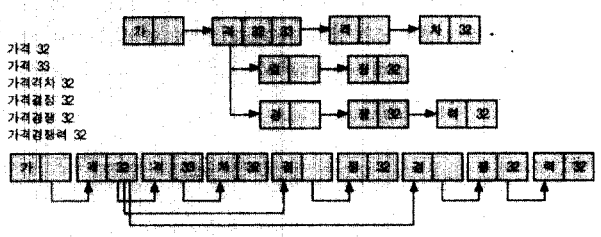
본 시스템에서 개발된 사전의 구조는 한국어 사전 표제어의 전자적 저장 방법으로 가장 널리 활용되고 있는 TRIE에 기반한다. (그림 3)는 TRIE 사전의 노드 구성을 나타낸다.



(그림 3) TRIE 사전 음절 노드의 구조

TRIE 사전의 한 노드는 총 6byte로 구성된다. 처음 2byte는 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 10bit를 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 정보 값이 저장되게 된다.

(그림 4)은 사전 표제어가 실제로 전자사전에 저장되는 구조를 나타낸다. (그림 4)의 상단 부분은 표제어가 전자사전에 저장되는 의미적인 구조를 나타내고 아래 부분은 실제로 메모리와 디스크에 적재되는 구조를 나타낸다. 한 표제어가 여러 개의 사전 정보를 포함하는 표제어의 경우는 각 표제어 정보를 나타내는 음절 노드를 연속해서 저장한다. 이때 다수의 표제어 정보를 저장하기 위해서 6바이트음절 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.



(그림 4) TRIE 사전에 표제어가 저장되는 구조

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

<표 5>는 사전 탐색 관련 모듈을 설명한 것이다. 크게 일반사전 탐색, 조사사전 탐색, 어미사전 탐색 함수로 나뉜다. 각 함수가 캡슐화되어 있으므로 전달되는 매개변수가 비교적 많은 특징이 있다.

〈표 5〉 사전 탐색 관련 모듈 API

함 수 명	프로토타입	기 능
SearchDic	DWORD SearchDic(HANGUL * h_word, UWORD h_word_len, UWORD h_idx, DIC_RESULT * result, UWORD * res_idx) ;	일반사전 탐색
SearchJosaDic	DWORD SearchJosaDic(HANGUL * h_word, UWORD h_word_len, UWORD first_jong_only, JEDIC_RESULT * result, UWORD * res_idx) ;	조사사전 탐색
SearchEomiDic	DWORD SearchEomiDic(HANGUL * h_word, UWORD h_word_len, UWORD first_jong_only, JEDIC_RESULT * result, UWORD * res_idx) ;	어미사전 탐색
매개변수설명	• h_word : 조합형 입력어절 • h_word_len : 입력어절 길이 • h_idx : 첫글자의 인덱스 • result : 사전 탐색 결과 저장 버퍼	

4.2 어절 형태소 분석

한국어의 어절 구성 형태에 대한 일반적인 규칙은 각 모듈별로 쉽게 구현이 가능하다. 이 논문에서 개발된 시스템

은 최장 일치에 의한 좌-우 분석 기법을 사용하였으므로 우선적으로 실질 형태소에 대한 사전 탐색 후에 모든 세부 분석 절차가 이어지게 된다.

입력 어절에 대한 형태소 분석은 크게 (1) 체언 분석, (2) 용언 분석, (3) 부사류 분석, (4) 독립언 분석, (5) 미등록어 분석 (6) 수사 분석 등으로 나뉜다. 이 각각의 분석 단계는 하나하나의 모듈로 구현되어 순서대로 이루어진다. 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, 위의 <표 6>은 단어형성규칙에 따른 세부 어절 분석 기능을 열거하고 있다.

〈표 6〉 단어생성규칙에 따른 어절 분석 기능

분류	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-", "차-", "키-"
	"이다" 처리	
	복합명사 처리	
용언 분석	[illegible]	
	[illegible]	
	[illegible]	
	[illegible]	
	불규칙 활용형 처리	[illegible]
	어미 분석	[illegible]
부사 분석	보조사 사전 탐색	
미등록어 분석	[illegible]	
	형식 형태소 분석	
수사 분석	수사패턴검사, 단위 명사 분석	
	후미어 형태소 분석	

각 분석 분류에 따른 세부 분석 항목이 위에서 설명한 각 모듈의 세부 모듈로 구성된다. 특정 분석 분류에서 분석 순서는 대부분 고정되어 있고 몇가지 예외처리를 거치게 된다. 예를 들어, 체언 분석에서 용언화 접사 처리 모듈이 성공하게 되면 그 다음으로 이어지는 모듈은 용언 분석 분류에서의 어미 분석이 된다. 이것은 상당한 문맥 이동이다. 이러한 문맥 이동이 손쉽게 이루어지는 가장 중요한 이유는 어미 분석 세부 모듈이 완벽하게 캡슐화되어 있기 때문에 그 결과로 용언화 접사 처리 부분에서 간단한 모듈 호출로 쉽게 문맥의 이동이 가능하기 때문이다.

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.

<표 8>은 어절 분석시에 가장 먼저 호출되고 모든 하부 분석 모듈의 출발점이 되는 어절 분석 함수를 보여주고 있다. 사전 탐색이 완료된 어절에 대해서 탐색된 사전 정보에 따라서 품사별로 분석 방법이 달라지게 된다.

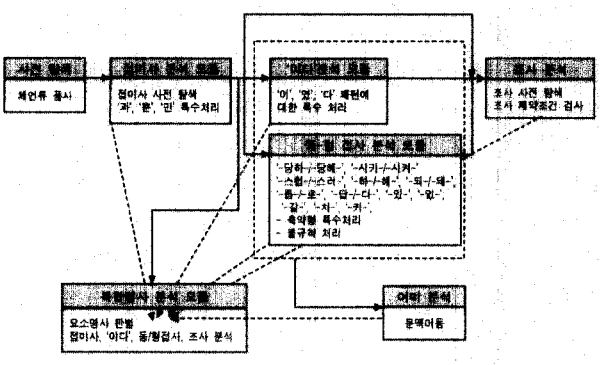
〈표 8〉 어절 분석 주 모듈 API

함 수 명	프로토타입	기 능
KAnalysisWord	DWORD KAnalysisWord(HANGUL *h_word, UWORD h_word_len, DIC_RESULT *dic_result, UWORD dic_res_idx, DWORD mode) ;	어절 분석

4.2.1 체언 분석 모듈

체언은 실질형태소와 형식형태소 사이의 변이가 적고 형식형태소의 종류가 많지 않으므로 비교적 다른 품사에 비해서 분석이 쉽다. 그러나 분석 단계에 따라서 많은 규칙이 포함되어야 하고 때에 따라서는 용언 분석 문맥으로의 문맥 이동이 발생하므로 이에 대한 적절한 구현 및 처리가 필요하다. 또한 정보 검색 분야에서 보면 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 다양한 방법론이 적용되어야 한다. (그림 5)는 사전에서 체언으로 판명된 좌측 최장 부분 어절의 나머지 부분 어절을 분석하는 과정을 도식화한 것이다. 실선으로 표시된 화살표는 모듈의 호출 순서 및 문맥 흐름을 나타내고 점선으로 표시된 화살표는 복합명사 모듈이 사용하는 각각의 체언 분석 모듈에 대한 호출을 도식화한 것이다. 복합명사 분석은 각 단위명사 분석 사이에 일반적인 체언 분석 기능을 사용해야 한다. 만일 체언 분석 문맥을 이루는 각 단위 분석이 모듈화되어 있지 않다면 복합명사용 체언분석 모듈들을 다시 구성하거나 기능이 제한된 모듈만을 사용할 수 밖에 없다. (그림 5)에서 각 세부 기능들을 구성하는 요소들이 모두 모듈화되어 있으므로 이를 사용하는 복합명사 분석 모듈을 구성하기가 쉽다. 이러한 특성을 이용하여 재귀호출을 사용하지 않고도 528라인 정도의 소규모 루틴으

로 복합명사 분석 모듈을 구성할 수 있었다.



(그림 6) 체언 분석 및 호출 경로

사전 탐색에서 명사로 판별된 어절에 대해서 접미사 분석을 한 후, "이다" 조사가 붙거나 동사/형용사화 접사가 붙은 어절에 대해서 어미분석을 수행하였다. 또한 각 단계별로 제약조건을 두어서 각 분석 단계에 들어갈 필요가 없는 부분 어절에 대해서는 바로 다음 단계로 넘어 갈 수 있도록 하였다. 조사사전 탐색 부분은 조사 바로앞 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 어절에대하여 복합명사 분석을 수행한다. 본 시스템에서는 다음과 같은 두 가지의 가정 하에서 복합명사 분석을 수행하였다.

<표 9>는 체언 분석에 사용되는 기능 모듈들을 열거한 것이다. 마찬가지로 분석 기능별로 모두 완전히 모듈화가 되어 있다.

〈표 9〉 체언 분석 모듈 API

함수 프로토타입	기 능
DWORD CheckIda(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"이다" 검사
DWORD CheckJosa(HANGUL * h_word, UWORD j_pos, UWORD jong) ;	조사 제약조건
DWORD VAJubSaProc(HANGUL * h_word, UWORD h_word_len, UWORD info) ;	동-형 접사검사
DWORD CheckDabDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"답다" 검사
DWORD CheckRobDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"롭다" 검사
DWORD CheckSiKiDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"시키다" 검사
DWORD CheckDangHaDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"당하다" 검사
DWORD CheckDoiDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"되다" 검사
DWORD CheckHaDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"하다" 검사
DWORD CheckSeuRubDa(HANGUL * h_word, UWORD h_word_len, UWORD jong) ;	"스럽다" 검사
DWORD CheckCompNoun(HANGUL * h_word, UWORD h_word_len, HANGUL pre_char, UWORD info) ;	복합명사 분석

어 있다.

복합명사 분석 모듈은 입력 매개변수로 이전 분석 위치 바로 이전에 위치한 음절을 입력으로 받아서 음절간 제약조건 검사가 가능하도록 하였다. 본 시스템에서는 복합명사 분석시에 두 가지 가정을 세우고 이 가정에 따라서 복합명사 분석을 수행한다.

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

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

(가정 1)은 접미사를 실질형태소로 보는 것이 아니라 형식형태소로 간주하는 것이다. 접미사에는 한 글자로 구성된 단어들이 많으며 이를 복합명사의 일부분으로 분석하기 위해서는 많은 제약조건에 따른 오류가 발생하기 쉽다. 따라서 복합명사를 구성하는 접미사가 붙은 명사는 하나의 단일명사로 간주하여 사전에 포함시키는 것이 오류를 최소화시키는 방법이다. (가정 2)는 (가정 1)과 의미적으로 연관되는 가정이다. 한 글자로 구성된 명사를 복합명사의 요소명사로 포함시키면 복합명사 분석 자체가 무의미할 수 있다. 따라서 두 글자 이상으로 구성된 명사만이 복합명사의 요소명사가 될 수 있다고 가정한다.

이와 같은 가정 하에서 입력 어절에 대한 복합명사의 분석은 다음의 (그림 6)과 같은 알고리즘으로 수행된다.

```
Ssub : 입력 어절에서의 현재 분석 대상 부분 어절
Ui : 분석 대상 음절
Ssub = Ui Ui+1 Ui+2 … Ui+k
for ( j = 0 to k-1 ) {
    if ( j == 1 ) continue ;
    Ui+j 위치에서 사전 탐색 수행 ;
    if (2음절 명사)
      ( VERTEX[Index]에 현재 사전 탐색 결과 저장 ; )
}
연속된 명사 리스트를 찾아서 DADJ에 인덱스를 저장 ;
/* 후보 단위명사 그래프 traversal */
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 DADJ 에서
            추출하여 스택에 저장 ;
        접미사 검사 ; "이다" 조사 검사 ;
      용언화 접사 검사 ; 조사 검사 ;
    }
}
```

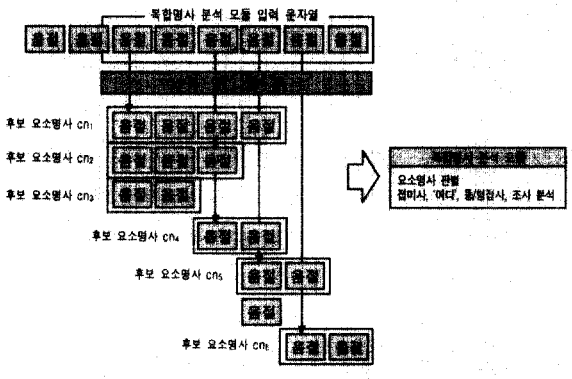
(그림 6) 복합명사 분석 알고리즘

이 분석 알고리즘의 가장 큰 특징은 입력 단어에 대해서 복합명사의 요소명사일 가능성이 있는 후보 단위명사를 추출하여 이들을 하나의 그래프로 구성한다는 것이다. 각 그

래프의 입력 edge와 출력 edge는 각각 현재 후보 단위명사의 이전 후보 단위명사, 이후 후보 단위명사의 vertex를 가리킨다. 실제 복합명사 분석은 이 그래프가 구성된 후에 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 7)은 위의 알고리즘을 이용하여 복합명사가 분석되는 개념과 구조를 나타낸다.



(그림 7) 복합명사 분석

(그림 7)에서 입력 문자열의 두 번째 음절위치를 제외한 모든 위치에서의 사전 탐색 결과로 여러 개의 후보 요소명사가 추출된다. 이 후보 요소명사를 그래프의 한 vertex로 간주하고 그래프의 각 노드를 스택을 이용하여 방문하면서 복합명사 분석 모듈을 호출하게 된다. 만일 입력 문자열의 끝까지 분석이 성공되면 하나의 복합명사 분석 결과로 분석 결과 버퍼에 저장하게 된다.

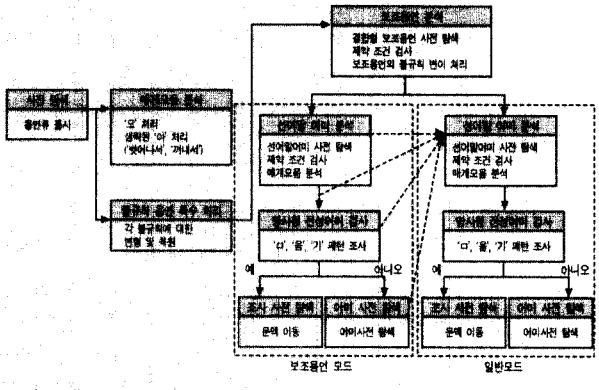
본 논문에서 수행된 복합 명사 분석 알고리즘은 모든 분석 대상 음절 위치에서의 사전 탐색에 따른 사전 탐색 횟수의 증가에 대한 문제점이 있을 수 있으나, 일반적인 복합 명사 분석 알고리즘에서 사용하고 있는 재귀적 호출이나 복잡한 모듈의 구현을 피할 수 있는 장점이 있다. 만일 분석 상의 오류나 추가로 수행되어야 할 기능적 모듈을 전체 복합 명사 모듈에 추가시킬 때는 후보 요소명사 그래프 생성 부분은 수정할 필요없이 그래프의 각 노드들을 방문하며 실제 복합명사를 분석하는 부분만을 수정하면 된다.

4.2.2. 용언 분석 모듈

형태소 분석 단계에서 가장 복잡한 부분이 용언 분석 모듈이다. 용언 분석이 복잡한 이유는 실질형태소와 형식형태소 사이에서 다양한 형태의 변이가 생기기 때문이다. 따라서 이러한 변화 형태를 규칙화하고 규칙화하기에 까다로운 복잡한 부분은 실제 루틴으로 구현함으로써 전체적인 분석 모듈이 복잡해진다.

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 논문에서는 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.

용언 분석은 크게 두 가지 모드로 수행될 수 있다. 보조용언 모드와 일반 모드가 그것이다. 보조용언 모드는 보조용언으로 시작하는 형식형태소를 분석하는 모드이고, 일반 모드는 선어말 어미에서부터 형식형태소가 시작하는 모드이다. 모드의 분기는 보조용언 분석 모듈 내의 보조용언 사전 탐색 기능에서 수행된다. 각 모듈의 수행 절차를 하나로 묶어서 분석하기보다는 모드별로 다른 모듈을 구성하여 두 모듈을 유기적으로 결합하도록 하였다. 다시 말해서, 보조용언 모드에서 분석 중에 분석에 실패하게 되면 바로 일반 모드의 시작점으로 분기하는 형태이다. 일반 모드는 그 분석 절차가 기본적인 용언의 형식형태소에 준하므로 분석 실패에 대한 분기가 필요가 없다. 단순히 용언 분석 실패를 윗부분에 알려주면 된다.



(그림 8) 분석 모드에 따른 용언 분석

(그림 8)은 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 어미 처리부의 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 루트

로 넘어가서 조사 사전 탐색을 수행하게 된다.

<표 10>에 나온 함수들은 용언 분석에 사용되는 대표적인 함수들이다. 'CheckEomi'는 위의 (그림 8)의 알고리즘대로 용언형 형식형태소에 대한 형태소 분석을 수행한다. 또한 각 불규칙 형태별로 분석 함수를 분리하고 원형 복원과 함께 제약조건 검사를 수행하게 된다.

〈표 10〉 용언 분석 모듈

함수 프로토타입	기 능
DWORD S_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len)	"ㅅ"불규칙
DWORD D_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len) ;	"ㄷ"불규칙
DWORD B_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len, …	"ㅂ"불규칙
DWORD H_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len, …	"ㅎ"불규칙
DWORD RUE_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len) ;	"르"불규칙
DWORD EU_IrrProc(HANGUL * h_word, UWORD h_word_len, UWORD pre_morp_len, …	"으"불규칙
DWORD CheckEomi(HANGUL * h_word, UWORD h_word_len, WORD first_jong_only) ;	어미 분석
DWORD IsFirstPreEomiChar(HANGUL one_char) ;	선어말어미 첫글자

4.2.3 부사, 독립어, 관형사 분석

이 논문에서 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.

4.2.4 수사 형태소 분석

이 논문에서 개발된 수사 형태소 분석 모듈은 모든 수사 패턴들을 처리하기 위하여 일반적인 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.

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

<표 8>은 수사의 기능과 형태에 따른 수사 표제어 정보를 나타낸다. 단위명사 결합형 수사는 사용빈도가 아주 높은 수사와 단위명사의 결합형을 사전에 추가함으로써 분석 시 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.

〈표 11〉 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)

이 사전 정보를 바탕으로 약 91개의 수사 표제어와 263개의 단위명사를 정제하여 사전으로 구성하였으며, 수사 형태소 분석을 수행하기 위하여 (그림 9)과 같은 분석 알고리즘이 적용되었다.

```
if (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 9) 수사 형태소 분석 알고리즘

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 가정하고 명사 다음의 조사나 접미사를 분석하듯이 수사의 후절어를 분석해야 한다. 수사 다음의 명사 분석(SUSA_NOUN 모드)도 마찬가지이다. 이와 같이 분석 모드에 따른 다양한 어절 분석 API를 제공함으로써 보다 세부적인 수사 형태소 분석을 수행할 수 있게 된다.

〈표 12〉 수사 분석 모듈 API

함수 프로토타입	기 능
DWORD CheckSusaWord(HANGUL * hword, DWORD hword_len, UBYTE * word, tMORP_RESULT * sMorpResult, UWORD * sMorpResult_Index) ;	수사 검사 메인 모듈
DWORD SusaSearch(UBYTE * word, DWORD * dic_ret, DWORD * dic_ret_idx) ;	수사 패턴 탐색 모듈
DWORD SusaUnitSearch(HANGUL * hword, DWORD hword_len, UWORD * idx) ;	단위 명사 탐색 모듈
DWORD CheckSusa(UBYTE * word, HANGUL * hword, DWORD hword_len, DWORD mode, DWORD * dres, DWORD dres_idx, tMORP_RESULT * sMorpResult, UWORD * sMorpResult_Index) ;	수사 검사 세부 분석 모듈

4.2.5 미등록어 분석

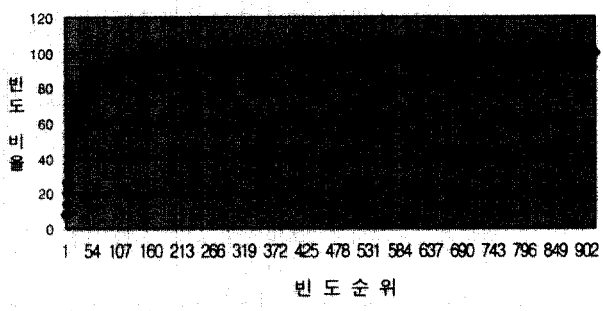
형태소 분석 수행 후에 분석에 실패한 어절들은 미등록어 분석을 거치게 된다. 본 논문에서 개발된 미등록어 분석기는 일반적으로 사용되는 역방향 분석에 의한 형식형태소 분리 기법이 아닌 형식형태소 사전을 통한 단순 사전 탐색에 의해서 수행된다. 형식형태소를 추출하기 위해서 전체 9,438,209어절로 구성된 말뭉치에 대해서 본 논문에서 개발된 형태소 분석기를 이용하여 형태소 분석을 수행하고 각각 체언, 용언 뒤에 붙는 형식형태소를 따로 분리 추출하였다. 또한 실질형태소와 형식형태소의 경계 음절 바이그램을 추출하여 통계치에 적용하였다. 추출된 형식형태소 사전 및 음절 바이그램에 대한 분석 정보는 다음 표와 같다.

〈표 13〉 형식형태소 분석 정보

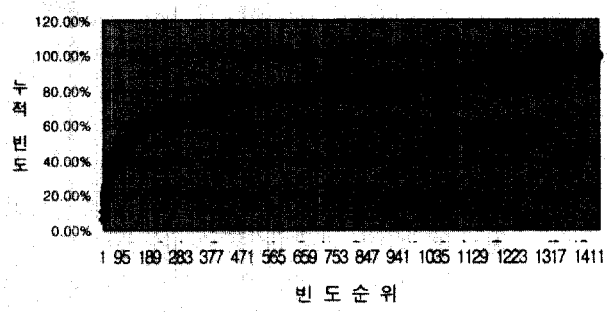
	개 수	단일화된 개수	빈도가 1인 형태소 개수
체언형 형식형태소	7,241,254	6,377	702
용언형 형식형태소	4,116,926	66,000	19,262
경계 음절 바이그램	7,241,524	46,384	2,692

체언형 형식형태소에 비해서 용언형 형식형태소와 경계 바이그램의 종류가 매우 많음을 알 수 있다. 이는 용언형 형식형태소는 불규칙 활용이나 탈락, 그리고 축약 등에 의해서 실질형태소와 형식형태소의 경계 부분에 변이가 많이 생기기 때문이다.

(그림 9)와 (그림 10)은 각각 체언형, 용언형 형식 형태소에 대한 빈도 순위 변화에 따른 누적 빈도를 나타낸 것이다. 체언형 형식 형태소는 빈도 순위 100등까지의 누적 분포 비율이 전체 형태소 개수의 90% 이상을 차지하는 반면, 용언형 형식형태소는 50%정도만 차지하고 있다. 이는 앞에서도 설명하였듯이 체언형 형식형태소에 비해서 규칙 및 불규칙 활용이 많이 발생하여 형식형태소의 종류가 광범위하게 적용되는데 그 이유가 있다.



(그림 9) 체언형 형식형태소의 빈도 순위별 누적빈도



(그림 10) 용언형 형식형태소의 빈도 순위별 누적빈도

대용량 말뭉치에 대한 형태소 분석을 수행할 때, 본 논문에서는 협소 문맥(local context)을 이용한 HMM 품사 태거를 사용한다. 이 품사 태거에서 추출된 형태소 분석 결과에서 확률값으로 상위 3위 안에 나온 형태소 분석 결과를 저장하게 된다.

위와 같이 구성된 사전을 이용하여 입력어절에 대한 형식형태소 사전 탐색 및 실질-형식 형태소 경계 음절 바이그램 사전 탐색을 수행한 후에 체언 및 용언 판별을 수행하고 판별된 결과에 따라서 형태소 분리 작업을 수행한다. 본 논문에서 개발된 시스템에서는 미등록어는 체언일 가능성이 더 높다고 가정하고 우선 체언형 미등록어인지를 검사한다. 체언형 미등록어 판정 확률은 다음과 같이 계산된다.

$$P_{unk} = \lambda_1 \times \frac{Freq_{bTail}}{TotalFreg_{bTail}} + \lambda_2 \times \frac{Freq_{nTail}}{TotalFreg_{nTail}} + \lambda_3 \times \frac{TailLen}{WordLen}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

첫 번째 항은 실질-형식 형태소 경계 바이그램 빈도이고 두 번째 항은 체언형 형식형태소 빈도이며 세 번째 항은 형식형태소의 길이에 전체 어절 길이를 나눈 값이다. 이렇게 계산된 확률값이 특정 임계치를 넘으면 용언형 형식 형태소에 대한 검사는 하지 않는다.

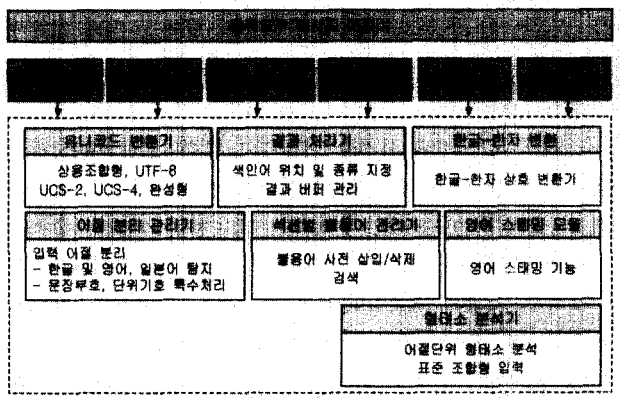
〈표 13〉 활용꼴 형태소 복원 내용

구 분	복 원 내 용
체언형	조사 제약 규칙 검사
	조사의 첫음절이 될 수 없는 음절 검사
	체언의 마지막 음절이 될 수 없는 음절 검사
	휴리스틱스 규칙 검사
용언형	"어/아/었/았"에 대한 "ㅅ"불규칙 검사 및 복원
	"어/아/었/았"에 대한 "ㄷ"불규칙 검사 및 복원
	"ㄹ"탈락 활용형태 복원
	"와→오아", "워→우어", "여→이어", "왜→외어"
	기타 불규칙 검사

체언으로 판별된 미등록어에 대한 형태소 분리는 비교적 쉬운 반면에 용언으로 판별된 미등록어에 대한 형태소 분리는 불규칙 활용 및 축약꼴이 발생할 수 있으므로 이에 대한 원형 복원 작업을 수행하게 된다.

5. 형태소 분석 시스템의 응용

본 논문에서 개발된 한국어 형태소 분석 시스템은 정보검색 시스템인 K-2000에 적용되었다. 범용 정보검색 시스템은 다양한 색인 형태를 제공할 수 있으므로 형태소 분석 시스템 위에 IDX 모듈이라고 불리는 색인 모듈을 탑재하여 전체 시스템에 적용되었다. IDX 모듈의 구조는 다음 그림과 같다.



(그림 12) IDX 구조

유니코드 변환기는 모든 종류의 유니코드와 조합형 및 완성형 코드간의 변환 기능을 수행한다. K-2000은 기본적으로 UTF-8을 저장 기준 코드로 정하고 있다. 따라서 색인 결과도 UTF-8로 변환되어 제공된다. 형태소 분석기 내부에도 어절 분리기가 내장되어 있으나, 시스템의 효율적인 결합을 위해서 어절 분리 및 관리기를 IDX로 끌어올렸다. 불용어 관리 및 영어 스태밍 모듈도어절 분리 및 관리기와 함께 유기적으로 동작하게 된다. 한자 변환 색인이나 영어 스태밍 여부 등은 IDX의 옵션으로 지정하여 사용자가 상황에 맞게 적절히 색인 기능을 수행하도록 하였다. 현재 IDX에는 총 6가지의 색인 유형이 제공된다. 각 색인 유형에 대한 설명은 다음 표에 나타나있다.

〈표 14〉 색인 유형

색인 유형	색인 예
	정보검색에 관한 연구
IndexByMA	[정보], [검색], [연구]
IndexByToken	[정보검색에], [관한], [연구]
IndexAsIs	[정보검색에 관한 연구]
IndexAsIsMA	[정보검색에 관한 연구], [정보], [검색], [연구]
IndexByChar	[정], [보], [검], [색], [에], [관], [한], [연], [구]
IndexAsNumeric	[123.45], [10000]

형태소 분석기를 색인 기능의 근간으로 두고 사용자의 요구사항에 부합하는 다양한 색인 형태를 제공함으로써 시스템의 효용성을 극대화하였다. 현재 IDX는 K-2000의 주 색인모듈로서 탑재되어 구동되고 있다.

개발된 IDX 시스템에 대한 속도 측정 실험을 수행하였다. 실험 환경 및 실험 데이터는 다음의 <표 15>와 같다.

〈표 15〉 속도 측정 실험 환경 및 대상 데이터

분석 환경	O/S : 와우 리눅스 CPU : Pentium III 933 * 2 M/M : 1 Gbyte
대상 자료	한국일보 데이터(2,160,897 토큰)

색인 유형은 가장 기본적인 유형인 IndexByMA로 지정하였으며 색인 시간에는 유니코드 변환, 어절 분리, 형태소 분석, 결과 저장 시간 등이 모두 포함되었다. 다음 <표 16>은 속도 실험 결과이다.

〈표 16〉 속도 측정 결과

	추출 색인어 개수	분석 시간	1초당 분석 어절
전체 결과	4,602,934(어절)	126.46(초)	17,087(어절/초)
형태소 분석	-	89.48(초)	24,149(어절/초)

위 실험 결과에서 대상 데이터의 토큰 수보다 추출된 색인어 개수가 더 많은 이유는 태깅을 수행하지 않고 모든 형태소 분석 결과를 색인어로 제시하기 때문이다. 또한 유니코드 변환이나 어절 분리 및 결과 저장에 걸리는 시간이 전체 색인 시간의 약 30%를 차지하는 것을 알 수 있다.

6. 결론 및 향후 연구방향

본 논문에서 구현된 시스템은 어절 분석 속도를 높일 수 있는 최적의 알고리즘으로 구현되었으며, 모듈화된 하부 시스템의 유기적이고 효율적인 결합에 중점을 두고 각 모듈별 성능 및 속도 검증이 가능하도록 하였다. 또한, 재귀적 복합명사 분석을 탈피하여 시스템 부하를 줄였으며 다층적 수사 패턴 인식에 기반한 수사 형태소 분즘은 분석 가능한 범위 한도 내에서 거의 100%의 정확도를 나타낸다. 이 논문에서 구현된 시스템도 마찬가지이다. 따라서 시스템의 정확도나 검증에 초점을 맞추기보다는 시스템이 어떻게 구성되어 있느냐에 중점을 두어 논조를 기술했다.

본 논문에서 개발된 시스템의 특징은 어절 분석 속도를 높이기 위하여 품사 사전의 구조화와 탐색 방법에 대한 다양한 접근 방법의 평가를 통해 최적 기능을 추가하였다. 형태소 분석기의 특징상 성능에 대한 실험 및 평가는 오해의 소지가 상당히 많다. 이 논문에서 개발된 시스템은 품사 태거가 포함된 자동 색인 시스템이 아니므로 분석 정확도에 대한 실험 결과는 그 의미가 매우 제한적일 수밖에 없다. 따라서 불명확한 정확도에 대한 실험결과를 제시하기보다는 시스템의 기능을 분석적으로 설명하는데 중점을 둠으로써 전체적인 시스템의 언어 일반성에 대한 효율적인 처리 능력과 언어 특수성에 대한 집중화된 방어능력에 대해서 설명하였다.

실제로 구현된 시스템에 대한 성능 평가 실험을 여러 차례 수행하였고, 결과에 대한 다양한 검증을 시도하였으나 형태소 분석 시스템의 특징상 그 성능을 객관적으로 평가하기에는 무리가 있었다. 그 이유는 다음과 같다. 형태소 분석 시스템은 크게 분석 모듈 구현과 사전 구성으로 구성된다. 어떻게 보면 정확도를 판가름하는 대부분의 분석 기능이 사전에 기초를 두고 있다고 해도 과언이 아니다. 이러한 상황에서 모듈에 대한 정확도의 판가름에 대한 실험은 전체 시스템 평가에 중요한 요소(factor)가 되지 못한다. 또한 형태소 분석 시스템은 분석 가능한 모든 분석 결과를 제시한다. 일반적으로 대부분의 분석 알고리즘의 알고리즘을 구현하였으며, 전체적인 시스템 구조를 디자인함에 있어서 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.

본 논문에서 구현된 형태소 분석 시스템이 정보검색 시스템과 결합하여 검색 결과의 정확도와 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 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 철자오류 수정 시스템을 개별적으로 개발하여 성능을 검증하였으나, 이를 유기적으로 통합하는 작업은 향후의 연구로 남겨 두었다.

참 고 문 헌

[1] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위논문, 1993.

[2] 최성필, "오류분석정보와 복합명사의 의미처리규칙 및 말뭉치를 이용한 철자 교정기의 성능 개선", 부산대학교 전자계산학과 석사학위논문, 1998.

[3] 한경수, 이도길, 임해창, "통합 정보 검색을 위한 과학기술문서 색인 및 요약 시스템의 개발", 제5회 한국 과학기술 정보인프라 워크샵 논문집, 2000.

[4] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기", 부산대학교 전자계산학과 석사학위논문, 1995.

[5] 채영숙, 김재원, 김민정, 권혁철, 한국어 철자 검색을 위한 형태소 분석 기법, "'91 우리말 정보화 잔치", 국어 정보학회. pp. 179-186, 1991.

[6] 채영숙, "연어 규칙에 기반한 한국어 문서 교정시스템의 구현", 부산대학교 전자계산학과 박사학위논문, 1998.

[7] 강승식, "다층 형태론과 한국어 형태소 분석 모델", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.140-145, 1994.

[8] 강승식, "음절 특성을 이용한 한국어 불규칙 활용 어절의 형태소 분석 방법", 1993년도 제5회 한글 및 한국어 정보처리 학술발표논문집, 1993.

[9] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 부산대학교 전자계산학과 박사학위논문, 1997.

[10] 동아 새국어사전, 서울 : 동아출판사, 1995.

[11] 이영식, "사전 근사탐색과 Heuristics를 이용한 한국어 철자 오류 교정 시스템 구현", 부산대학교 전자계산학과 석사학위논문, 1994.

[12] 강승식, 권혁일, 김동렬, "한국어 자동 색인을 위한 형태소 분석의 기능", 한국정보과학회 춘계 학술 발표논문집, 제22권 제1호, pp.929-932, 1995.

[13] 심준혁, 김준석, 이근배, "통계와 규칙을 이용한 강인한 품사 태거", 제1회 형태소 분석기 및 품사태거 평가 워크샵 논문집, pp.60-75, 1999.

[14] 이윤재, 김선배, 김길연, 최기선, "모듈화된 형태소 분석기의 구현", 제1회 형태소 분석기 및 품사태거 평가 워크샵 논문집, pp.123-136, 1999.

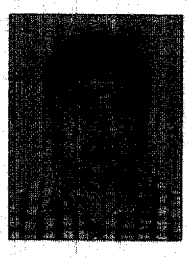
[15] 장동현, 맹성현, "학습데이터를 이용하여 생성한 규칙과 사전을 이용한 명사추출기", 제1회 형태소 분석기 및 품사태거 평

가 워크샵, pp.13-22, 1999.
[16] 최재혁, "형태소 분석을 통한 한영 자동 색인어 추출 시스템", 정보과학회논문지, 제23권 제12호, pp.1279-1288, 1996.
[17] 김태희, 박혁로, 신중호, "검색/요약/필터링을 위한 텍스트 이해 모형 연구", 제3회 소프트과학 워크샵 논문집, 1999.
[18] 이근용, 박기선, 이용석, "Two-level 한국어 형태소 해석에서의 복합명사 처리", 2002 정보과학회 봄 학술발표논문집, pp. 505-507, 2002.
[19] 김남철, 서영훈, "형태소 분석기 CBKMA와 색인어 추출기 CBKMA/IX", 제1회 형태소 분석기 및 품사 태거 평가 워크샵 논문집, pp.50-59, 1999.
[20] 강승식, "한국어 수사어절의 유형 분류 및 정규화", 한국정보과학회 추계학술발표논문집, pp.127-189, 1999.
[21] 김수남, 원상현, 권혁철, 주종철, 이상기, "의미정보를 이용한 한국어 복합명사 분석", 한국정보과학회 추계학술발표논문집, pp.195-197, 1999.
[22] 신기철, 신용철, "새우리말 큰사전", 삼성출판사, 1994.
[23] Baeza-Yates, Ricardo, and Ribeiro-Neto, Berthier, "Mordern Information Retrieval, New York : ACM Press," 1999.
[24] Cahill, L. J., "Syllable-based Morphology," Proceedings of the 13th International Conference on Computational Linguistics, Vol.3, pp.48-53, 1990.
[25] Charniak, Eugene, "Statistical Langual Learning," A Bradford Book, Cambridge : The MIT Press, 1993.
[26] Kang, S. S., "A Statistical Approach to Syllable-based Morphological Analysis," Proceedings of the International Conference on Computer Processing of Chinese and Oriental Language, 1992.
[27] Kelly, Douglas G., Introduction to Probability. London : Macmillan Publishing Company, 1994.
[28] Koskcnniemi, K., "Two-level Model for Morphological Analysis," Proceedings of the 8th International Joint Conference on Artificial Intelligence, pp.683-685, 1983.
[29] Kwon, H. C., Chae, Y. S. and Jeong, G. O., "A Dictionary-based Morphological Analysis," Proceedings of NLPRS '91, pp.87-91, 1991.
[30] Kwon, H. C. and Karttunen, L., "Incremental Construction of a Lexical Transducer for Korean," Proceedings of the 15-th International Conference on Computational Linguistics, Vol.2, pp.1262-1266, 1994.
[31] Manning, Christopher D. and Hinrich Schutze, Foundations of Statistical Natural Language Processing, Cambridge : The MIT Press, 1999.

<부 록>

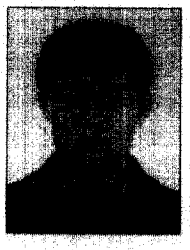
〈표 2-1〉 분석 사전 정보

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



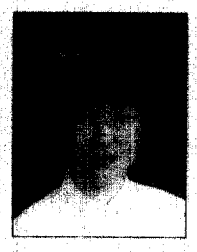
최 성 필

e-mail : spchoi@kisti.re.kr
1996년 부산대학교 전자계산학과 졸업 (학사)
1998년 부산대학교 대학원 전자계산학과 졸업(석사)
1998년~2000년 연구개발정보센터 연구원
2001년~현재 한국과학기술정보연구원 연구원
관심분야 : 자연어처리, 정보검색, 기계학습, 데이터마이닝 등



서 정 현

e-mail : jerry@kisti.re.kr
1987년 한양대학교 수학과 졸업(학사)
1987년~1994년 시스템공학연구소 연구원
1994년~2000년 연구개발정보센터 연구원
2001년~현재 한국과학기술정보연구원 연구원
관심분야 : 정보검색, 자연어처리 등



채 영 숙

e-mail: yschae@ysu.ac.kr
1980년 부산대학교 전산통계학과(학사)
1991년 부산대학교 계산통계학과(석사)
1998년 부산대학교 전자계산학과(박사)
1995년~1999년 한국전자통신연구원 연구원
1999년~2000년 한국과학기술원 연구원
2001년~현재 영산대학교 멀티미디어공학부 전임강사
관심분야 : 게임공학, NLP, AI, HCI