좌우접속정보를 이용한 명사추출기

안동언
전북대학교 전자정보공학부
duan@moak.chonbuk.ac.kr

A Noun Extractor using Connectivity Information

Dong Un An
Division of Electronic and Information Engineering, Chonbuk National University

요 약

본 논문의 명사추출기는 정보검색시스템을 위한 색인어 추출기로 좌우접속정보를 이용한 형태소해석을 통하여 얻어진 형태소들 중에서 명사를 추출한다. 본 형태소해석기는 형태소해석을 위한 언어지식과 어절 분리 엔진을 분리하여 수정과 확장이 용이하게 하였다. 사용한 언어지식은 좌우접속정보로서 한 어절을 이루는 형태소들의 품사간의 접속여부를 행렬로 표현한 것이다. 어절 분리 엔진은 사전을 참조하여 한 어절에서 최장일치법에 의해 형태소를 분리하고 좌우접속정보를 참조하여 형태소 분리가 올바른지를 판단한다. 형태소들의 품사분류는 표준 태그셋을 기반으로 음절 정보를 추가하여 확장하였다. 형태소를 해석한 결과 미등록어가 발생하였을 때 미등록어에서 명사를 추정하는 모듈이 없기 때문에 재현율은 좋지 않았다.

1. 서론

본 논문의 명사추출기는 정보검색시스템을 위한 색인어 추출기로 좌우접속정보를 이용한 형태소해석을 통하여 얻어진 형태소들 중에서 명사를 추출한다. 본 형태소해석기는 형태소해석을 위한 언어지식과 어절 분리 엔진을 분리하여 수정과 확장이 용이하게 하였다. 사용한 언어지식은 좌우접속정보로서 한 어절을 이루는 형태소들의 품사간의 접속여부를 행렬로 표현한 것이다. 어절 분리 엔진은 사전을 참조하여 한 어절에서 최장일치법에 의해 형태소를 분리하고 좌우접속정보를 참조하여 형태소 분리가 올바른지를 판단한다.

형태소들의 품사분류는 표준 태그셋을 기반으로 음절 정보를 추가하여 확장하였다. 또한, 명사 추출에 큰 영향을 미치지 않는 부사, 관형사, 조사, 어미 등의 세부 분류는 사용하지 않았으며 대분류만을 사용하였다.

사전은 10만 단어로 이루어져 있으며 명사가 8만 단어이다. 기능어인 조사와 어미는 복합형을 사전에 모두 수록하여 형태소해석 과정의 부담을 경감시켰다.

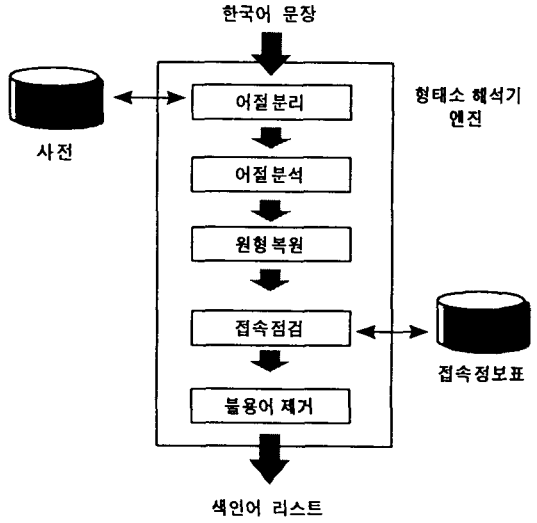
본 명사추출기에는 형태소를 해석한 결과 미등록어가 발생하였을 때 미등록어에서 명사를 추정하는 모듈이 없다.

2. 명사 추출기의 구성

명사추출기는 <그림 1>에서 보는 바와 같이 크게 형태소 해석기 엔진, 사전, 접속정보표의 세 부분으로 구성되

어 있다.

형태소 해석기 엔진은 한국어 문장을 입력으로 받아서 어절 단위로 형태소 해석을 하여 색인어 리스트를 출력으로 내 보낸다. 어절 분리, 어절 분석, 원형 복원, 접속 점검, 불용어 제거 등의 5단계로 구성되어 있다.



<그림 1> 명사추출기 구성

사전은 형태소와 그 형태소에 해당하는 품사 태그를 가지고 있다. 사전 색인 구조로는 자소 단위의 트라이를 사용하며 구현에 있어서는 배열을 이용한다. 사전에 있는 등록어의 총 수효는 101,303개이다.

접속정보표는 어절을 구성하고 있는 형태소들 간의 결합 관계를 정의한 것으로 단위 어절 문법이라고 할 수 있다. 분할된 형태소들 간의 결합 제약 조건을 접속정보표에 기술하고 각 형태소마다 결합 정보를 사전에 기술함으로써 분할된 형태소간의 결합 타당성을 검사한다.

각 구성 요소에 대하여 다음 각 절에서 자세히 설명한다.

3. 형태소 해석기

언어학에서의 형태소 해석이란, 각각의 단어와 어절의 문법적 결합관계를 밝히기 위해 의미의 최소 단위인 형태소로 해석하는 것을 가리킨다. 반면에 한국어 정보처리를 위한 형태소 해석은 어절에 대한 언어적 정보를 합성하기 위하여 어절을 구성하는 사전 표제어 혹은 형태소를 형태소 해석 문법에 기반하여 파악하는 것을 말한다. 따라서 한국어 정보처리를 위한 형태소 해석은 주어진 입력문장으로부터 최소 의미 단위인 형태소를 사전에 저장된 색인어와 형태소 해석 방법을 이용하여 추출하는 것이다.

한국어 형태소 해석 시에 고려해야 하는 한국어의 형태론적 변형의 원인에는 음운론적 이형태, 준말, 모음 탈락, 모음 축약, 모음 조화, 매개모음 삽입, 불규칙 활용 등 다양하다. 또한, 이러한 형태론적 변형이 연쇄하는 형태소들의 경계에서 발생하고 형태소가 분리되는 위치를 인식하기가 어려우므로 가능한 모든 음절 경계에서 형태소를 분리하면서 형태론적 변형을 검사하여야 한다. 이러한 사실이 한국어 형태소 해석을 어렵게 한다. 또한 이러한 형태론적 변형이 일어나는 위치는 실질 형태소와 형식 형태소가 연쇄할 때이다. 한국어 형식 형태소는 조사와 어미로 실질 형태소와는 달리 폐쇄군(close set)이므로 사전에 모두 등록하면 되지만 형식 형태소들의 결합에 의한 복합 형식 형태소도 매우 발달되어 있어서 그렇게 간단한 일은 아니다.

이러한 문제들을 해결하기 위하여 다양한 한국어 형태소 해석 기법이 제시되어져 왔다. 본 명사추출기에서는 접속정보를 이용한 최장일치법을 채택하기로 한다. 형태론적 변형을 처리하는 형태소 해석 알고리즘은 규칙을 기반으로 하며 형태소 해석 방향은 bottom-up 방식이다. 어절의 검색 방향은 좌측에서 우측 방향으로 하며 사전을 트라이(trie) 구조로 구성하면 사전을 탐색하는 방향과 일치시킬 수 있어서 탐색의 효율을 높일 수 있다. 형식 형태소는 최소 단위인 단위 형태소뿐만 아니라 결합이 가능한 모든 유형의 복합 형식 형태소를 하나의 단위로 사전에 수록하여 형태소 분할의 정확성과 속도를 높인다. 형태소 처리 단위로는 자소 단위로 하여 형태론적 변형 현상을 정확하게 처리하도록 한다.

형태소 해석기의 처리 과정은 어절 분리, 어절 분석, 원형 복원, 접속 점검, 불용어 제거의 5단계로 이루어져 있다.

3.1 어절 분리

어절 분리 단계에서는 입력되는 문자열이 한글만으로 구성된 것이 아니므로, 먼저 어절 타입을 검사한다. 어절 타입은 한글 문자열, 영문 문자열, 숫자, 특수 문자로 구분하며, 시스템 사전에 엔트리로 존재하는 영문+한글(NBA농구단, CD플레이어 등), 영문+숫자(A4, M16 등)의 형태를 복합 형태로 정의한다. 이러한 어절 타입 검사의 정보를 가지고 한글에 대해서 코드 변환을 실시한다. 원형 복원 단계에서 형태론적 변형 현상을 처리할 수 있도록 내부 코드로는 3byte 조합형 한글 코드를 사용한다.

3.2 어절 분석

어절 분리가 끝나면 실제로 사전을 탐색하여 어절에

포함된 가능한 모든 형태소들을 분리해낸다. 형태소 해석 알고리즘은 규칙을 기반으로 하며 형태소 해석 방향은 bottom-up 방식이다. 어절의 검색 방향은 좌측에서 우측 방향으로 하며 사전을 트라이(trie) 구조로 구성하면 사전을 탐색하는 방향과 일치시킬 수 있어서 탐색의 효율을 높일 수 있다.

3.3 원형 복원

원형 복원 단계는 불규칙 활용 어절로부터 어간과 어미의 원형을 복원하는 과정을 말하며, 처리를 위해 기본적으로 불규칙 활용의 정의 및 분류가 필요하다.

불규칙 활용이란 어간과 어미가 결합할 때 어간의 모양이 달라지거나, 어미가 예외적인 형식으로 결합하는 것으로 정의되는데, 일반적으로 언어학에서는 모든 용언에 대해서 변화가 일어나는 자동적 교체와 일부 용언에 대해서만 변화가 일어나는 비자동적 교체로 구분하며, 전산학에서는 처리의 편리성을 위해서 어미의 유형이나, 어간과 어미의 변화 여부에 따른 분류가 시도되었다.

본 색인어 추출기에서는 먼저 어간과 어미의 원형 및 어미가 변이된 형태 즉, 불규칙 활용 형태를 사전에 수록하고 해석시 원형을 추론해내는 방법을 사용한다. 시스템 사전에 존재하는 12가지의 불규칙 활용은 ㅅ불규칙동사-형용사, ㄷ불규칙동사, ㅂ불규칙동사-형용사, 르불규칙동사-형용사, 우불규칙동사, 여불규칙동사-형용사, 러불규칙동사-형용사, ㅎ불규칙형용사, 거라불규칙동사, 너라불규칙동사, 으탈락동사-형용사, ㄹ탈락동사-형용사 등이다.

3.4 접속 점검

접속 점검 단계에서는 사전에서 찾아진 형태소간의 결합 타당성을 검사하여 그 결과를 반환한다. 접속정보표는 형태소들간의 접속정보를 갖고 있어 분할된 형태소들간에 서로 접속이 가능한지를 검사하기 위한 테이블이다. 어절에서 앞에 있는 형태소의 우접속 정보와 뒤에 있는 형태소의 좌접속 정보간의 접속 가능 여부를 표시하고 있다. 따라서 접속 점검 부분에서는 접속정보표를 참조하여 형태소들간의 결합 타당성 여부를 결정한다. 접속정보에 대해서는 5장에서 설명한다.

접속정보표의 구조는 [표 1]과 같다.

[표 1] 접속정보표

		뒤 형태소의 좌접속정보						
		NCP	...	JC_N	JC_YL	JC_NL	JC_Y	JC_C
앞 형태소의 우접속 정보	NCP_N	O	...	O	X	O	X	O
	NCP_Y	O	...	X	O	X	O	O
	NCP_L	O	...	X	O	O	X	O
	...	...	...	...	...	...	...	...
	J	X	...	O	O	O	O	O

3.5 불용어 제거

형태소 분석이 모두 끝난 후 형태소들 중에서 색인으로서 가치가 없는 불용어를 제거한다. 불용어는 사전에 등재되어 있다. 현재 6,500여 개의 불용어가 사전에 들어 있다.

이번 MATEC99에서는 불용어를 제거하지 않았다.

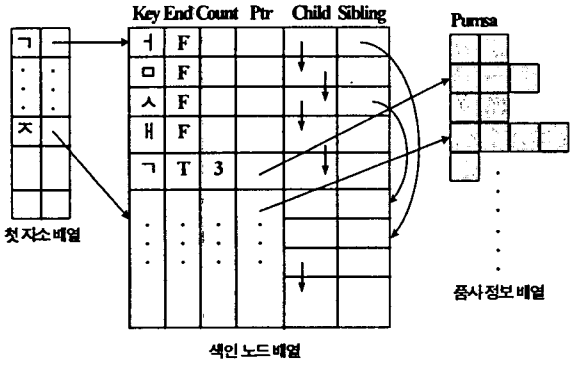
4. 사전

한국어 형태소 해석기에서는 사전 내 등록어를 검색하기 위하여 빠른 검색 속도를 제공하는 효율적인 사전 색인 구조가 요구된다. 기존의 시스템들에서 사용하는 색인 구조로는 B+-트리, 해쉬, 자소별 트라이, 음절별 트라이 등이 있으며, 이 중 B+-트리와 해쉬는 트라이 구조의 사전에 비해 사전 검색 횟수가 많아 한국어 형태소 해석을 위한 사전 색인 구조로 적합하지 않다. 본 명사추출기에서 사용하는 사전에서는 사전 색인 구조로 자소 단위의 트라이(Trie)를 사용하며, 구현에 있어서는 배열을 이용한다. 트라이는 키 값 전체가 아니라 그 일부에 의해 각 레벨의 분기가 결정되어지는 차수가 2이상인 트리 구조를 말하는데 가변적 크기의 키 값을 다룰 때 특히 유용한 색인 구조이다.

트라이를 이루는 색인 노드는 세 개의 자료 필드와 세 개의 링크 필드로 구성된다. 첫 번째 자료 필드에는 분기를 위한 키에 해당하는 한글 한 자소를 저장하고, 두 번째 자료 필드에는 끝 글자 정보를 저장한다. 세 번째 자료 필드에는 끝 글자일 경우에 엔트리가 가지는 품사의 개수를 표시한다. 네 번째 필드는 품사 정보 배열을 가리키는 포인터이며, 다섯 번째 필드는 자식 노드(child node)를 가리키며 여섯 번째 필드는 형제 노드(sibling node)를 가리킨다.

트라이는 키워드 트라이와 사전 트라이 두 가지 형태로 구분하여 사용한다. 키워드 트라이는 색인 작업 시 색인어들의 문헌 빈도값을 구하기 위해서 임시로 색인어들을 저장할 경우 사용되는 형태이며, 사전 트라이는 시스템 사전을 트라이로 구성할 경우 쓰이는 형태이다.

<그림 2>는 검색이라는 엔트리가 트라이로 입력된 한 예이다. 텍스트 사전은 엔트리, 품사, 품사 ....의 형태로 구성되어 있으며, 텍스트 사전은 프로그램 수행할 때 초기에 주기억장치에 적재되어 자소 단위의 트라이 인덱스 구조로 변환된다.



<그림 2> 자소별 트라이 사전 구조

링크드 리스트를 이용한 트라이 구조를 사용하는 사전들은 노드를 삽입할 때마다 매번 주 기억 장치를 할당하고, 삭제할 때마다 노드에 사용된 주 기억 장치를 시스템에 반환해야 한다. 또한 생성된 트라이 구조를 디스크에 색인 파일로 저장할 때 노드의 수 만큼 링크를 따라가며 다른 노드를 가리키는 포인터가 그 노드의 디스크에서의 적당한 주소를 가리키도록 변환시켜 주어야 하며, 로드할 때에는 노드마다 주 기억 장치를 할당하여 디스크에서의 포인터가 가리키는 노드를 메모리에서도 가리키도록 그 값을 적절히 변환하여야 한다.

트라이를 구성하는 기본 자료 구조로서 각 원소가 링크드 리스트의 한 노드에 해당하는 배열을 사용한다. 노드의 링크 필드는 다른 노드의 주소를 갖는 것이 아니라 다른 노드의 배열에서의 인덱스를 갖는다. 프로그램 수행 초기에 배열을 위한 주기억장치를 한꺼번에 할당받고, 배열의 관리를 위하여 현재까지 사용되지 않은 노드들을 가리키는 free_node를 두어 관리한다. 만약 새로운 노드가 추가될 때 더 이상 비어 있는 노드가 없을 경우 지금까지 사용되던 배열보다 큰 새로운 배열을 할당하고 이 새로운 배열에 지금까지 사용되던 배열의 내용을 복사한 후 사용되던 배열은 시스템에 반환한다. realloc함수를 이용한다.

자소별 트라이 구조에서 사용하지 않는 필드들로 인한 주기억공간의 낭비를 없애기 위해 배열을 유형별로 나누어 분리시키는 것이 필요하며 사전 검색 속도의 향상을 위해 첫 자소 배열에 적절한 수의 링크 필드를 두는 것이 필요하다.

사전에 있는 등록어의 총 수효는 101,303개이다. 품사별 등록어의 수효는 [표 2]와 같다.

[표 2] 품사별 사전 등록어의 수효

품사	수효	품사	수효	품사	수효
보통명사	69,408	고유명사	12,084	의존명사	385
대명사	179	수사	99	동사	2,098
형용사	736	관형사	124	부사	1,412
감탄사	143	조사	4,715	어미	3,361
불용어	6,558				

조사와 어미의 경우에는 복합형을 사전에 모두 수록하여 형태소해석 과정의 부담을 조금 경감하였다. 그렇지만, 어미의 경우에 존칭과 시제를 나타내는 선어말어미는 생산성이 높기 때문에 독립된 형태로 사전에 수록하여 복합어미의 수효가 너무 많아지는 것을 방지하였다.

동사와 형용사의 경우에는 실행 사전을 만들면서 불규칙 용언의 변형도 집어넣어서 접속정보점검을 쉽게 하였다. 사전에 수록된 우리말 용언 중에서 불규칙 용언은 1,500개 정도 된다.

5. 접속정보

한국어에서 형태소들은 어절을 구성하기 위하여 다른 형태소들과 결합한다. 이와 같이 어절을 구성하고 있는 형태소들 간의 결합 관계를 정의한 것이 접속 정보로서 단위 어절 문법이라고 할 수 있다. 분할된 형태소들 간의 결합 제약 조건을 접속정보표에 기술하고 각 형태소마다 결합 정보를 사전에 기술함으로써 분할된 형태소간의 결합 타당성을 검사할 수 있다.

접속정보표를 구성하기 위해서는 품사 체계가 필요하다. 국어학에서 품사 분류는 학자마다 다를 정도로 다양하다. 국어학의 분류는 한국어 정보처리에서 그대로 사용하기에 부적합하다. 국어학에서는 언어학적인 관점에서 형태와 기능에 따라 분류를 하지만, 한국어 정보처리에서는 정보처리라는 측면에서 품사를 분류해야 하므로 국어학의 품사 분류와는 다르다. 또한 한국어 정보처리에서는 기계인 컴퓨터가 처리해야 하므로 품사를 좀 더 자세하게 분류해야 한다. 이 품사 체계는 형태론적 품사 체계로 형태소 해석 관점에서 애매성을 줄이고 정확한 형태소 분석 결과를 제공하여야 한다.

본 명사추출기에서 사용하는 품사체계는 국어공학센터의 국어정보베이스를 위한 한국어 품사 태그를 기반으로 한다.

품사 분류가 방대해짐에 따라 어절 네트워크의 표현이 어려워지게 되므로, 어절 네트워크의 표현 방식을 변경한다. 좌접속 정보와 우접속 정보를 분리한다. 형태소의 좌우접속정보는 파생법과 합성법에 근거한 단어의 형성

을 표현하는 방법의 일종이다. 이 방법의 기본 개념은 형태소를 두 개의 독립적인 분류 체계에 따라 나누고 두 분류 체계간의 접속성을 조사하여 이를 형태소간의 접속 모델로 하자는 것이다. 좌접속 정보는 좌측에 붙을 수 있는 형태소를 기준으로 분류한다. 따라서 조사의 경우 명사의 형태에 따라 세분된다. 우접속 정보는 우측에 붙을 수 있는 형태소를 기준으로 분류한다. 따라서 체언의 경우 조사의 이형태에 따라 세분된다. 즉, 좌접속 정보는 접속적 성격을 나타내고 우접속 정보는 문법적 성격을 나타낸다.

명사와 조사와의 접속관계를 통해서 접속정보의 예를 들어본다. 명사가 조사와 결합을 할 때 명사의 제일 뒤 글자의 종성에 따라서 결합하는 조사 이형태가 다르다.

(예1) 학교 + 를
집 + 을

즉, 유종성과 무종성에 따라서 결합하는 조사 이형태가 다르다. 그런데 "로"와 "으로"의 경우에는 앞에 오는 명사의 제일 뒤 글자의 종성이 "ㄹ"인 경우에 유종성이지만 무종성과 결합하는 "로"와 결합한다.

(예2) 학교 + 로
칼 + 로
집 + 으로

따라서, 명사의 경우에 "무종성(NCP_N)", "유종성(NCP_Y)", "ㄹ종성(NCP_L)"로 구분한다. 조사의 경우에도 종성 종류에 따라 결합하는 종류가 다르기 때문에 명사의 분류에 맞추어서 "무종성과 결합하는 조사(JC_N)", "ㄹ 종성을 제외한 유종성과 결합하는 조사(JC_YL)", "ㄹ 종성을 포함한 무종성과 결합하는 조사(JC_NL)", "유종성과 결합하는 조사(JC_Y)", "모든 조사와 결합하는 조사(JC_C)"로 구분한다. "JC_C"는 이형태가 없는 조사를 의미한다.

어미의 경우에도 양성모음과 음성모음을 구분하여야 하지만 명사의 추출이 목적이었기 때문에 이러한 구분을 하지 않았다.

명사 추출에 큰 영향을 미치지 않는 부사, 관형사, 조사, 어미 등의 세부 분류는 사용하지 않았으며 대분류만을 사용하였다.

품사간의 접속정보뿐만 아니라 이형태에 따른 접속정보도 품사 분류에 반영함으로써 형태소해석 엔진에서 접속여부를 점검하는데 단순히 접속정보표만을 찾아보면 된다.

6. 결론

본 명사추출기의 가장 큰 장점은 형태소해석을 위한 언어지식과 어절 분리 엔진을 분리하여 수정과 확장이 용이하다는 것이다. 사용한 언어지식은 좌우접속정보로서 한 어절을 이루는 형태소들의 품사간의 접속여부를 행렬로 표현한 것으로 음절정보를 품사분류에 추가하여 확장함으로써 어절 분리 엔진의 동작을 단순하게 하였다. 따라서 엔진의 핵심이 되는 프로그램은 매우 적었으며 속도는 빨랐다.

본 논문의 명사추출기는 정보검색시스템을 위한 색인어 추출기이기 때문에 완벽한 형태소해석을 하지는 않는다. 용언에 대한 처리는 어간과 어미의 변화에 대한 자세한 분석을 하지 않고 용언이라는 것을 확인만 하였다. 또한, 명사 추출에 큰 영향을 미치지 않는 부사, 관형사, 조사, 어미 등의 세부 분류는 사용하지 않았으며 대분류만을 사용하였다.

형태소 해석은 최장일치법을 사용하여 형태소를 찾았으며 이 방법에 있어서 사전 검색을 쉽게 하도록 사전은 트라이로 구성하였다. 사전은 10만 단어로 이루어져 있으며 명사가 8만 단어이다. 기능어인 조사와 어미는 복합형을 사전에 모두 수록하여 형태소해석 과정의 부담을 경감시켰다.

본 명사추출기에는 형태소를 해석한 결과 미등록어가 발생하였을 때 미등록어에서 명사를 추정하는 모듈이 없다. 따라서, MATEC99의 평가에서 재현률이 좋지 않았다.

이번 MATEC99 평가를 위해서 기존에 작성된 명사추출기의 수정에 노력을 기울이지 않았다. 명사 추정 모듈을 첨가하고 사전을 정련하여 향상된 결과를 얻어야 할 것이다.

참고문헌

[1] 신동욱, 안동언, MIDAS 기반 정보검색 시스템 개발, 최종보고서, 한국전자통신연구소, 1996

[2] 신동욱, 장재우, 안동언, IR과 DBMS의 효과적인 통합, 최종보고서, 한국전자통신연구소, 1997

[3] 안동언, 품사 사전 규칙과 시범 패키지, 국어정보처리 기술 개발 제3차년도 최종보고서, 한국과학기술원, 1997

[4] 안동언, 확장 품사 사전 규칙과 보급 패키지, 국어정보처리기술 개발 제3차년도 최종보고서, 한국과학기술원, 1997

[5] 안동언, 국어 형태 통사 태그의 표준화, 1997년도 제2회 우리말 정보처리 규격 심포지움, 한국과학기술원

인공지능연구센터, 고려대학교, 1997, pp.51-59
[6] 안동언, 한국어 태그 집합, 전자공학회지, 제24권, 제9호, 1997, pp.1030-1037
[7] 이영주, 자동색인을 위한 한국어 형태소 분석 알고리즘, 1989년도 제1회 한글 및 한국어 정보처리 학술발표 논문집, 1989, pp.240-246
[8] 조영환, 차희준, 김길창, 확장 사전 환경에서의 한국어 형태소 형태와 생성, 1993년도 제5회 한글 및 한국어 정보처리 학술발표 논문집, 1993, pp.355-368
[9] 최기선, 한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구, 첨단요소기술 과제 제2차년도 최종보고서, 과학기술처, 1992
[10] 최기선, 남영준, 김진규, 한영균, 박석문, 김진수, 이춘택, 김덕봉, 김재훈, 최병진, 한국어정보베이스를 위한 형태.통사 태그 표준에 관한 연구, 인지과학 제7권, 제4호, 1996년