계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현

이 은철, 이 종혁

포항공대 전자계산학과

The Implementation of Korean Morphological Analyzer Using Hierarchical Symbolic Connectivity Information

Eun-Chul Lee, Jong-Hyeok Lee

Department of Computer Science, POSTECH

요 약

본 논문은 구문해석, 의미해석 등의 전처리 단계로서의 형태소 분석기 구현에 대해 기술하고자 한다. 먼저 기존의 접속정보의 단점을 보완하는 새로운 접속정보를 정의한다. 이 접속정보는 계층적구조를 가지고 심볼로써 표현되며, 기존의 좌우 두 가지 접속정보를 사용한 방법과는 달리 좌우를 하나로 통합한 정보를 사용한다. 따라서 접속정보 유지와 확장에 편의를 제공해 주고 접속정보 부여시 정확성을 더할 수 있고, 계층적구조를 살려서 접속정보표의 구성을 용이하게 한다. 또한 불규칙활용에 있어서는 사전정보에 의한 선언적 방법과 프러시져에 의한 절차적 방법의 장점을 살려 혼용하였다. 끝으로 앞에서 정의된 새로운 접속정보 방식의 장점을 살려서 정확한 분석 결과를 얻을 수 있는 형태소 분석기의 구현에 대해 설명한다.

I. 서론

최근 들어서 전산학 분야에 있어서 한글 사용이 날로 늘어가고 있는 실정이다. 또한 이러한 추세에 힘입어 다양한 한글 응용프로그램이 등장하고 있고 워드 프로세서의 기술발달로 인하여 더욱 한글에 관한 관심이 높아 가고 있다. 또한 한글을 처리하는 응용 프로그램인 철자 검색, 교정기, 띄어쓰기 검사/교정기, 정보검색기 등에 대해서도 많은 연구가 진행되어왔다[1]. 그런데 이러한 자연어(한국어)처리 응용에 기본이 되는 것은 형태소 분석이다. 특히 어미와 조사의 사용에 의해 단어의 형태적 변형(용언의 활용, 탈락, 불규칙...) 이 심한 한국어에서는 형태소 분석기의 역활이 중요하다[1]. 그런데 기존의 형태소 분석기를 의미해석, 구문해석 등의 상위레벨의 전단계로 사용하기에는 아직 부족한 형편이다. 그 이유는 형태소 응용프로그램[2,6] 에서는 부정확한 결과로도 어느 정도 만족한 결과를 얻을 수 있지만, 상위레벨의 처리를 위해서는 기본바탕이 되려면 보다 정확한 형태소 분석결과가 요구되기때문이다.

본 논문은 한국어처리와 관련하여 새로운 접속정보 방식을 이용하고 형태소 분석결과의 정확성에 중점을 둔 한국어 형태소 분석기의 구현에 대하여 설명한다.

II. 한국어 형태소 분석기에 대한 고찰

2.1. 기존의 연구들

형태소 분석이란 한국어 텍스트를 입력으로 하고 그것을 형태소 단위 - 사전의 표제어 단위- 로 분석하여 사전에 있는 정보(품사정보)와 함께 출력해 주는 것이다. 이러한 형태소 분석에서 형태소간의 접속 가능여부를 검사하는 방법으로는 접속정보를 이용한 방법과 접속규칙(filtering)를 이용한 방법 등이 있다. 이중에서 접속규칙을 이용하는 방법은 먼저 어절내에서 모든 형태소를 찾은 다음 그들을 어절규칙, 형태소간의 결합규칙을 통하여 형태소의 접속을 검사하여 가능한 결과만을 색출한다[8]. 이러한 방법은 접속정보를 사전에 형태소와 같이 부여하는 것이 아니라 단지 형태소의 품사정보만을 이용하기때문에 접속정보 방법에서의 어려움인 정확한 정보부여, 유지, 보수를 제거하였다. 그러나 또 한편으로는 접속정보 방법으로 형대소간의 접속을 제어하는 방법의 성질을 모두 결합규칙이라는 제약으로 만드는 것도 역시 어려운 문제가 된다. 접속정보 방법은 앞에서 설명한 어절규칙, 결합규칙 등과 형태소의 품사까지를 접속정보로 나타내어 이 접속정보를 사전에 표제어와 함께 등록하고 그것을 이용하여 형태소간의 접속을 제한하는 것이다. 이러한 방법은 형태소간의

접속을 제어하려면 품사에 대한 분류는 불론이고 더 세밀한 계층의 분류가 요구된다. 형태소간의 결합 여부를 나타내기 때문에 접속정보의 정확성 여부가 형태소 분석결과를 좌우한다[2,3]. 그런데 기존의 형태소 분석은 좌,우의 두 가지 접속정보를 사용하였기 때문에 접속정보를 부여하는데 있어서 정확성을 유지하기가 매우 어려웠다. 또한 접속정보의 부여시 고유한 숫자를 부여하였는데 이러한 경우에는 접속정보의 추가시 일관성을 갖기가 어렵다. 또한 숫자를 사용하는 방법으로는 그것이 무슨 품사인지를 파악하는 것이 힘들고 접속정보표 구성시 어렵게 된다. 그러므로 새로운 개선이 필요하다고 하겠다.

용언의 불규칙처리는 두 가지 방법이 행하여지고 있다. 프러시져에 의한 절차적(procedural) 방법과 용언의 사전 등록에 의한 선언적(declarative) 방법이 있다[5,6]. 이 중 첫째 방법은 용언의 기본형만을 사전에 등록하고 불규칙 활용의 어간은 절차적 방법을 통해서 찾는 것이다. 그리고 둘째 방법은 용언의 변형 형태까지 사전에 다 등록하고 형태소간의 접속조건을 기술해서 사전으로부터 그 기본형을 찾는 방법이다. 여기서 절차적 방법은 불규칙 활용이 일어나는 경우를 모두 조사하는 과정에서 불필요한 잘못된 처리를 거칠 수도 있다. 그리고 사전등록에 의한 방법은 용언의 불규칙 형태까지 표제어로 등록되어있으므로 불규칙 활용이 일어나지 않은 경우에도 용언의 변형형태를 사전에서 검색할 수 있다. 따라서 필요없는 사전검색을 통한 오버헤드(overhead)가 생길 수 있다. 효과적인 분석을 위해서는 이러한 상호간의 장단점을 서로 보완할 수 있는 방안이 필요하다.

2.2. 본 연구의 접근 방법

2.2.1. 접속정보.

앞에서 언급한 바와 같이 접속정보의 정확성이 형태소 분석의 정확성을 좌우한다. 접속정보란 형태소에 대하여 그 형태소가 갖고 있는 품사정보와 형태소의 접속을 제한하기 위해서 분류되어진 계층(hierarchy)에 대한 정보이다. 즉 자신의 품사를 나타내면서 또한 다른 형태소와의 접속을 위한 정보도 또한 지니고 있다. 이러한 정보를 이용하는 접속정보표에 대해서는 다음에 설명하기로 하겠다. 그러면 기존의 접속정보와 본 시스템의 접속정보를 비교하여 분석해보기로 하자.

1) 좌우 접속정보의 불필요성

좌,우의 두 가지 접속정보를 한 쌍으로 사용하였던 기존의 방식은 하나의 형태소가 좌측에서 볼때의 속성과 우측에서 볼때의 속성이

다르기 때문에 필요하였다. 그러나 이는 형태소분석만을 전제로 한다면 일본어의 경우에 적용되지 우리말의 경우는 필요치 않는다. 예를 들어 "고 수 있" 과 같은 형태소가 있다고 생각해보자. 그러면 이 형태소는 좌접속 정보로는 어미를 가지게 되고 우접속 정보로는 형용사의 어간과 비슷한 정보를 가지게 된다. 이와 같은 특별한 경우 때문에 좌우의 두 가지 정보가 필요했으나 어절을 입력단위로 하고 그 어절내에서의 분석을 목표로 하는 형태소 분석기에서는 위와 같은 형태소는 존재하지 않는다. 또한 존재한다고 하더라도 이러한 특별한 형태소에만 기존의 두 가지 접속정보를 부여하고 나머지 일반 형태소에는 하나의 접속정보를 부여하는 것만으로도 충분한 분석결과를 얻을 수 있다. 또한 접속정보의 유지, 보수에 있어서도 접속정보를 하나로 사용하는 것이 편리하다.

하나의 접속정보를 사용한다면 접속정보의 표현에 수정이 필요하다. 기존의 좌우접속정보의 특성을 살리기 위해서는 좌우정보를 통합하면서 또한 좌우정보의 고유한 특성을 나타낼 수 있어야 한다. 또한 접속정보 관리의 용이성을 위해 접속정보를 심볼로써 표현하고, 형태소간의 복잡한 접속관계의 효과적인 기술(description)을 위해서 접속정보를 계층구조로 구성하는 방법이 적당하다.

(2) 접속정보 표현방식의 비교

심볼 접속정보	●심볼에 대한 의미부여로 접속정보의 판독이 용이하다 ●심볼의 추가 등록을 통해 확장이 용이하다. ●계층적 구조를 갖는 접속정보로 인해 접속 정보표의 기술이 쉬워진다.
숫자 접속정보	● 이해하기가 어렵다. ●일관성 있는 번호 부여의 곤란으로 확장이 어렵다. ●접속정보표의 기술이 어렵다.

(표 1) 숫자를 이용하는 접속정보와 심볼을 이용하는 접속정보와의 비교

위에서 비교해 본 바와 같이 좌우 접속정보는 하나로 통합한 후 심볼로써 표현하고 계층적 구조를 갖도록 하는 것이 유리하다.

2.2.2. 불규칙 용언 처리.

불규칙 처리는 앞에서 언급한 두 가지의 방법을 혼용하였다. 일단 용언의 경우는 그 활용형까지 사전에 등록하는 방법을 택하였다. 그렇지만 용언의 간음화된 형태는 등록하지 않는다. 간음화의 처리는 절차적(procedural) 방식을 사용하였다. 절차적 방식을 사용할 경우는 미리 부여된 환경과 일치하면 처리를하게 되므로 불필요한 처리과정을 많이 거치게 된다. 이에 비해서 사전에 표제어를 등록하는 방식은 정보부여와 사전등록에 약간의 어려움이 있지만 불필요한 처리 과정을 거치지 않게 되므로 빈번히 사용되는 불규칙활용의 경우는 더 유리하다. 반면에 간음화된 형태는 그리 많은 빈도를 보이지 않기 때문에 사전의 표제어로써 등록하기 보다는 절차적 방법으로 처리하는 것이 용이하다. 하지만 이와 같은 방법을 사용할 경우는 불규칙 활용형인 형태소와 원형인 형태소가 접속할 수 있는 형태소에 제약을 주어야 하는 어려움이 있다. 그 어려움은 접속정보 부여를 위해서 계층분류를 심화함으로써 해결할 수 있다.

2.2.3 한글 코드.

한글코드는 완성형이 표준한글코드로 지정되었지만 완성형으로는 용언의 활용형 및 기타 음운현상을 처리하는 형태소 분석기에는 부적당하다. 따라서 본 시스템은 입,출력으로 한글 상용조합형을 사용하고 내부처리는 3바이트 코드를 써서 구현하였다. 조합형과 3바이트 코드간의 변형을 위한 과정은 분석기 내부에서 거치게 된다. 3바이트를 통해서 여러 음운현상을 쉽게 처리할 수 있다. 종성이 없는 경우도 3바이트로 처리하게 되므로 n바이트 형태보다는 메모리를 조금 더 쓰게 되지만 처리하는데 있어서 초성, 중성, 종성을 구분할 수 있으므로 편리하다

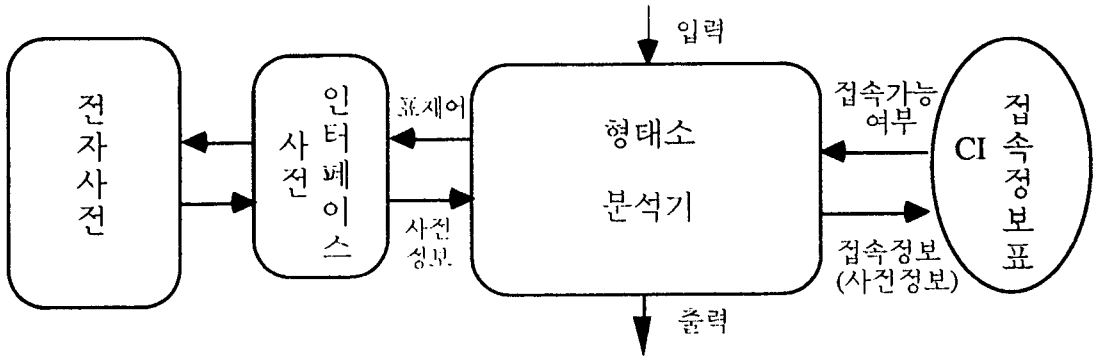
III. 형태소 분석기의 구현

3.1. 형태소 분석기의 목적

기존의 형태소 분석기는 주로 철자 교정기나 정보 검색기 등에 사용되었기 때문에 정확성이 더 중요시되었다. 본 시스템은 자연어처리의 관점에서 구문분석, 의미분석과 같은 상위레벨에서 사용하기 위한 전처리로서의 형태소 분석기를 목표로 한다. 따라서 보다 정확한 분석결과가 요구된다. 그 이유는 철자 교정기와 같은 경우 분석이 실패한 어절에 대해서만 교정을 시도하게 되므로 모호성이 있는 형태소 분석의 경우라도 임의의 분석이 성공하

기만 하면 된다. 그러나 그것이 오류를 포함한 분석결과라면 구문, 의미 해석을 위한 전처리 결과로서는 사용할 수가 없다. 그러므로 형태소 분석기는 정확성이 요구된다. 이를 위해서는 무엇보다도 접속정보의 정확성이 중요하다. 그리고 대다수의 철자 교정기나 검색기에서는 형태소 분석자체의 오류는 고려치 않고 있으나, 이러한 응용프로그램에도 더욱 정확한 결과를 얻을 수 있도록 하려면 역시 정밀한 형태소 분석기가 필요하게 된다.

3.2. 시스템 구성



[그림 1] 형태소 분석기의 전체구조도.

3.2.1 접속정보표

접속정보표는 형태소간의 접속여부를 나타내는 표이다. 이 표는 각 형태소들마다 부여된 접속정보를 이용하여 형태소 사이의 접속여부를 검사하는데 쓰인다. 본 분석기에서는 속도의 향상을 위하여 접속정보표를 메인메모리(main memory)에 띄워 놓는다. 이 때 하나의 심볼은 각기 하나의 계층을 나타낸다. 심볼에 있어서의 계층이란 그 심볼의 위치를 왼쪽부터 보았을 때의 순서를 의미한다. 접속정보표에는 다음과 같은 심볼을 사용한다.

? : 하나의 심볼을 대신한다. 예) MCN?m <-> MCNKm, MCNFm

* : 여러개의 심볼을 대신한다. 예) M*m <-> MCNKm, MDm

또한 접속정보 기술시, 심볼들의 계층적 구조에 의해 다음과 같이 상위 심볼로 모든 하위 심볼들을 표현할 수 있다.

M : M으로 시작하는 모든 심볼을 나타낸다.
(M*와 동일한 작용을 한다.)

MN : MN 으로 시작하는 모든 심볼을 나타낸다.
(MN 도 역시 M의 하위 계층이다.)

이러한 방법을 바탕으로 구성된 접속정보표의 형태는 다음과 같다.

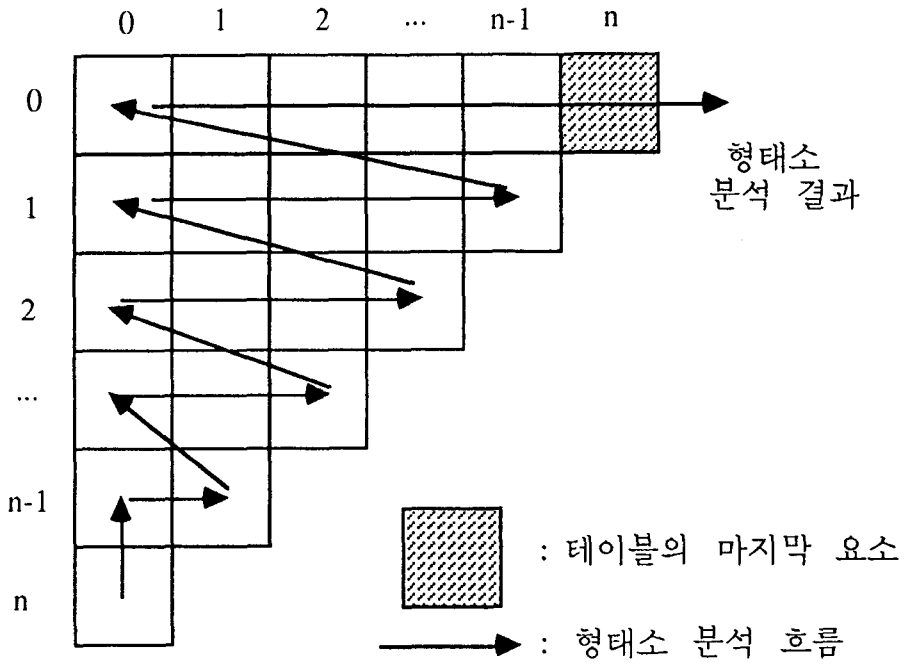
MN?C : MN?C MN?F +C
jRm : M*m D*m S*m mjNg mTe -?m
<예제 1> 접속정보표 구성의 예.

접속정보표 구성의 형태는 먼저 키가 되는 접속정보가 있고 그와 접속 가능한 정보들의 나열로 구성된다. 어절내에서의 두 형태소간의 접속을 검사할때 상대적으로 오른쪽에 있는 형태소의 접속정보가 키가 되고 접속정보표에서 그 키를 찾는다 그 후에 나머지 왼쪽 형태소의 접속정보가 접속정보표에서 앞서 찾은 키와 접속가능한 접속정보 목록에 속하는지 검사하여 그 접속여부를 결정한다. 여기서 "MN?C"의 심볼 하나하나는 앞에서 언급한 접속정보부여를 위한 계층을 나타낸다. 그리고 두 번째 줄의 접속정보를 보면 키가 되는 접속정보 "jRm"에 대해서 그와 접속이 가능한 정보들이 대개 심볼 "m"으로 끝나는 것을 볼 수 있다. 여기서 "m"의 의미는 무종성의 의미이다. 또한 "jRm"은 무종성 체언에 붙을 수 있는 격조사를 나타내는데 첫째 심볼이 "j"로 시작하는 접속정보는 모두 조사의 계층에 속한다. 그러므로 앞에서 언급했듯이 기존의 숫자를 부여하는 방법보다 더욱 일관성이 있고 그 의미파악에 있어서도 쉽다.

3.2.2 형태소분석 알고리즘

형태소분석에는 최장일치법, 최단일치법, head-tail기법, CYK 알고리즘을 이용한 tabular 파싱기법[3] 등 기존의 여러가지 방법이 있다. 이러한 기존의 알고리즘 중에서 형태소 분석기의 정확성을 검사하기 위하여 CYK알고리즘을 이용하기로 한다. 이를 통해서 가능한 모든 분석결과를 출력하여 그 정확성을 검사한다. 기존의 CYK 방식은 테이블 요소 하나하나를 모두 채워나가면서 분석했기 때문에 자소의 단위가 n개인 어절의 경우 $O(n^3)$의 시간 복잡도가 요구되었으나 본 연구에서는 알고리즘을 개선하여 $O(n^2)$의 시간 복잡도로 분석이 가능하다. 본 연구는 형태소 분석 처리에 있어서 n*n의 삼각형 테이블을 채워나가는 방식에 차이를 둔 것으로 테이블의 요소인 $n^2/2$ 만을 검사하여 마지막 요소를 통하여 가능한 모든 분석을 얻을 수 있게 된다. 그리고 분석시에 이전의 정보를 사용하여 처리한다면 더 빠른 분석을 가능

하게 할 것이다.



[그림 2]. CYK 알고리즘의 처리 과정

3.2.3. 전자 사전

전자사전에서 형태소에 대한 정보(접속정보)를 얻어오는 과정은 전자사전 인터페이스를 거친다. 이 때 전자사전 인터페이스에 상용조합형 코드로 구성된 표제어를 키로 하여 정보를 얻는다. 이러한 전자사전에서의 정보검색은 해쉬 함수(hash function)를 이용하여 속도의 향상을 꾀한다. 전자사전은 형태소와 그 형태소에 대한 접속정보로 구성된다. 접속정보는 앞에서 설명한 심볼 및 계층구조에 따라 사전 편집기를 통해서 표제어와 함께 온라인 방식이나 혹은 오프라인 방식으로 등록시킨 후 전자사전 인터페이스를 거쳐서 형태소 분석기에서 이용하게 된다.

3.3. 형태소 분석결과에 대한 해석

본 시스템은 C 언어를 이용하여 PC 상에서 상용조합코드를 입,출력으로 하는 형태소 분석기를 구현하였다. 입력화일은 역시 상용조합코드로 된 화일을 사용하여야 한다. 형태소 분석기는 입력단위가 어절이므로 입력화일을 다시 어절단위로 잘라서 분석한다. 그리고 한 어절의 분석결과는 분석 즉시

가능성있는 모든결과가 출력된다. 다음에 형태소 분석의 예를 보도록 하자.

입력 예 : 용기를 북돋웠다.
출력 예 :
어절 : 용기를
결과 1>> 용(품사 명사) 기(품사 접미사) 를(품사 조사)
결과 2>> 용(품사 명사) 기(품사 명사) 를(품사 조사)
결과 3>> 용기(품사 명사) 를(품사 조사)
어절 : 북돋웠다.
결과 1>> 북돋우(품사 동사) 었(품사 선어말어미) 다(품사 어말어미)

<예제 2> 형태소 분석결과의 예.

이 분석을 살펴보면 어절 "용기를"의 경우는 3 가지의 가능성있는 분석결과가 나왔다. 이러한 모호성은 형태소 단계에서는 분석이 불가능하다. 여기서 "기"라는 형태소의 경우에 품사가 접미사인 결과가 먼저 나온 이유는 그 형태소에 접속정보가 부여될 때 접미사를 나타내는 정보가 맨처음 부여되어 있기 때문이다. 형태소 분석시 하나의 결과만을 선택한다고 하면 접속정보 부여의 순서에서 빈도수가 높은 접속정보에 우선순위를 부여하면 된다. 다음으로 두 번째 어절 "북돋웠다"를 보면 이는 간음화현상이 일어난 경우이다. 사전에는 았/었 등의 선어말어미만을 등록해 놓고 앞에서 설명한 절차적인 방법에 의해서 "었"의 변형 형태와 동사의 어간을 찾아서 분석한다.

IV 결론 및 개선점

본 논문은 좌우 접속정보를 새롭게 하나의 심볼로 부여함으로써 기존의 방식보다 간편하고 확장성이 용이한 접속정보 방식을 제안하여 접속정보를 편리하게 이용할 수 있도록 하였다. 또한 이를 통해 보다 정확한 형태소 분석기를 구현하였다. 불규칙 처리에 있어서는 프러시져에 의한 절차적 방법과 변형형태의 사전등록에 의한 선언적 방법을 혼용함으로써 되도록이면 오버헤드를 줄이도록 시도하였다.

앞으로의 개선점으로는 미등록어 처리와 복합명사 분석에 있어서 휴리스틱한 처리의 추가를 들 수 있다. 이러한 처리를 위해서는 많은 입력자료의 분석을 통한 반복적인 자료 획득이 필요하다. 그리고 정확한 형태소

분석을 위해서는 꾸준한 접속정보의 보강이 필요하게 된다. 마지막으로 형태소 분석시 속도향상도 중요한 문제중의 하나이다.

V. 참고문헌

[1] 권혁철, "자연언어 처리 동향," 한국 정보과학회 인공지능연구회, 초청강연, 1991

[2] 송춘환, "한글 철자 및 띄어쓰기," 한국정보과학회 논문지, Vol.16, No.2, 1989.

[3] 김성용, "Tabular Parsing을 이용한 한국어 형태소 분석기" 한국과학기술원 석사학위논문, 1987.

[4] 김민정, "한국어 형태소 분석에서의 수사 처리," 한글 및 한국어 정보처리 학술발표논문집, 제 3회, 1991.

[5] 강승식, 김영택, "한국어 형태소 분석기에서 불규칙 용언의 분석 모형," 정보과학회 논문지, Vol.19, No.2, 1992.

[6] 강승식, 김영택, "한국어 형태소 분석기에서 선어말어미의 분석 모형," 정보과학회 논문지, Vol.18, No.5, 1991.

[7] 박영환, 김경서, 송만석, "말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현," 한국정보과학회 학술발표논문집, Vol18. No2., 1991.

[8] 여상화, 김용호, 이학주, 이정현, "다단계 필터링 능력을 갖는 형태소 분석기의 설계 및 구현," 한국정보과학회 학술발표논문집, Vol18. No2., 1991.