어절패턴 사전을 이용한 새로운 한국어 형태소 분석기

홍진표 ○ 차정원

창원대학교 컴퓨터공학과

virnmer@cwnu.ac.kr, jcha@cwnu.ac.kr

A New Korean Morphological Analyzer using Eojeol Pattern Dictionary

Jeen-pyo Hong○ Jeong-won Cha

Dept. of Computer Engineering, Changwon National University

요 약

본 연구에서는 어절패턴을 이용하는 새로운 방식의 한국어 형태소 분석기 KGuru-MA에 대해서 설명한다. KGuru-MA는 품사 부착 말뭉치에서 개방어를 생략하여 어절 패턴을 반자동으로 학습하여 어절 패턴 사전과 형태소 확률 정보 사전을 구성한 후, 이 사전을 이용하여 형태소를 분석한다. 본 형태소 분석기는 어절패턴을 사용하여 형태소 분석하기 때문에 기존 형태소 분석기에 존재하는 접속검사 과정이 생략된다. 또한, 형태소 분석 과정이 기존의 형태소 분석기에 비해 단순하여 기초 자연언어 처리 시스템이 가지는 강건성을 보장한다. 본 연구는 "21세기 세종기획 3차년도 말뭉치"를 이용한 실험 결과, 기존 형태소 분석기 못지 않은 성능을 보였다.

1. 서론

최근 인터넷의 발전으로 정보의 양이 방대하게 증가하였다. 이러한 현상은 컴퓨터를 이용한 정보의 효율적인 정리, 검색과 같은 자연 언어의 처리의 필요성을 크게 증대 시켰다. 이 가운데, 형태소 분석기는 자연 언어 처리의 가장 하위 단계로 상위 시스템의 필수적인 시스템 중 하나로 자리잡고 있다.

형태소 분석은 입력 문장에 대해 형태론적 변형과 형태소 분리 문제를 처리하는 과정으로 언어적 특성에 맞게 종속적인 형태로 구현이 된다. 특히, 한국어의 경우 영어와는 달리 조사와 어미의 의한 형태론적 변형이 생기기 때문에 보다 복잡한 구조를 취하고 있다.

지금까지 한국어 형태소 분석기에 대한 많은 연구가 이루어져 왔다. 그러나 이들 대부분이 외국 방법론을 한국어에 적용해 복잡하게 구현이 되어 있기 때문에 유지 및 보수가 매우 어려운 것이 현실이다.

본 연구는 형태소 분석기의 이러한 문제를 개선하고자 대량의 말뭉치로부터 자동으로 어절패턴을 이용한 형태소 분석 정보를 추출하여 이용할 수 있는 형태소 분석기 KGuru-MA를 개발하였다.

KGuru-MA는 기존의 형태소 분석기와 달리 접속검사에 해당하는 과정이 없기 때문에 형태소 분석기를 위한 접속 검사 사전을 구축할 필요가 없으며, 어절패턴 사전, 기호 사전, 형태소 확률 사전을 이용하는 것이 가장 큰 특징이다. 또한, 어절패턴 사전의 경우 사전의 내용이 직관적으로 구성되어 있어 사전 추가와 수정이 매우 용이하며, 패턴 정보를 이용한 모델로 형태소 분석기 KGuru-MA를 구축했다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 한국어 형태소 분석기 연구에 대해 정리하고 3장에서는 본 논문을 통해 구현한 어절패턴을 이용한 형태소 분석기에 대해 설명한다. 4장에서는 사전의 구축 방법을 알아보고 5장에서는 본 논문을 통해 구현한 형태소 분석기에 대한 실험을 진행하고 그 결과를 정리한다. 마지막 6장에서는 본 연구의 결론 및 앞으로의 연구 방향에 대해 설명하고 논문을 마치도록 한다.

2. 이전 연구

2.1 형태소 분석기 개발에 있어 고려할 점

형태소 분석기는 (1)강건해야 하고, (2)뛰어난 정확도를 요구하며, (3)미등록어에 대한 적절한 처리가 필요하며, (4)적절한 형태소 분석 후보를 결과로 내주어야 한다. 예를 들어, 인터넷 문서는 다양한 형태의 문서가 존재하는데 이들 문서는 입력 문자가 완벽하지 않은 것들이 많이 존재한다. 형태소 분석기는 이러한 문서를 적절하게 처리할 수 있도록 강건한 처리가 보장되어야 한다. 또한, 형태소 분석기는 자연언어 처리 시스템의 가장 첫 단계로 거쳐야 하는 작업이다. 이는 형태소 분석기의 정확도가 상위 시스템에 절대적인 영향을 미치게 되는데, 이러한 분석의 정확도를 높이기 위해 형태소 분석기는 재현률(recall)과 정확도(precision)가 모두 높아야 한다.

또한, 인터넷의 경우 하루에도 수 십여 개의 신조어들이 생성이 되는데 이들 단어는 형태소 분석기의 어휘 사전에는 없기 때문에 이들 미등록어에 대한 신뢰성 높은 결과를 내놓아야 한다. 그리고 형태소 분석 후보의 개수는 상위 시스템의 속도에 매우 높은 영향을 주기 때문에 되도록이면 후보의 개수는 적게 정확한 후보만을 내놓아야 한다.

2.2 형태소 분석 방법

형태소 분석 방법에 있어 지금까지 한국어 특성을 고려한 여러 연구들이 진행되어 왔으며, 기존에 한국어 형태소 분석에 있어 주로 사용되어 온 방법은 4가지 방법으로 요약할 수 있다.

첫번째로는 Two-level 형태론으로 이는 여러 개의 변형 규칙으로 구성되는 음운 현상을 하나의 두 단계 규칙으로 처리하도록 하는 방법이다[1]. 이 방법은 문자열 일치를 기반으로 하여 문법을 기술하는 문제 등 몇가지 문제가 있으나[2,3,4], 이를 한국어 특성에 맞게 수정하여 적용한 시도가 있었다[5,6].

다음으로는 Head-Tail 구분법으로 단어를 Head(어근)와 변형이 일어나는 Tail(문법 형태소)로 분리하여 결합 관계를 접속검사를 통해 분석하는 방법이다[7]. 그러나 이는 접속검사표 구성이 어렵다는 단점이 있다.

세번째로는 Tabluar Parsing 방법이다. 이 방법은 bottom-up 방식에 의한 형태소 분석 방법으로[8,9,10], Head-Tail 구분법과 마찬가지로 접속검사를 하기 때문에 접속 검사표를 구성해야 한다는 단점이 있지만 정확도가 높은 장점이 있다는 것이 특징이다.

네번째로는 최장일치법과 최단일치법으로 이는 가능한 모든 형태소로 분할해 형태소의 집합 중 가장 길거나 짧은 형태소를 포함하는 것을 우선적으로 검사하여 선택하는 방법이다[11].

그 외, 한국어의 음절의 통계적 특성을 반영하여 사전의 탐색 횟수를 줄여 속도 향상을 도모한 음절 정보를 이용하는 형태소 분석 방법[12]이 있다.

3. 어절패턴을 이용한 형태소 분석기

그림 1은 본 논문에서 제안한 어절패턴을 이용한 형태소 분석기의 전체 시스템에 대한 흐름도이다.

여기서, 어절패턴이란 어절에서 개방어(Open Word)의 정보를 제거하여 “~”로 표시하고 활용형은 변화된 것을 모두 포함하여 구성한 어절을 어절패턴이라고 한다. 예를 들어, “사람은”에서 “사람/명사+은/조사”에서 명사는 개방어이므로 어절패턴은 “~은”이 된다. 이러한 어절패턴에 대한 정보는 기본적으로 말뭉치로부터 반자동적으로 얻어내는데 이에 관한 내용은 4장에서 설명한다.

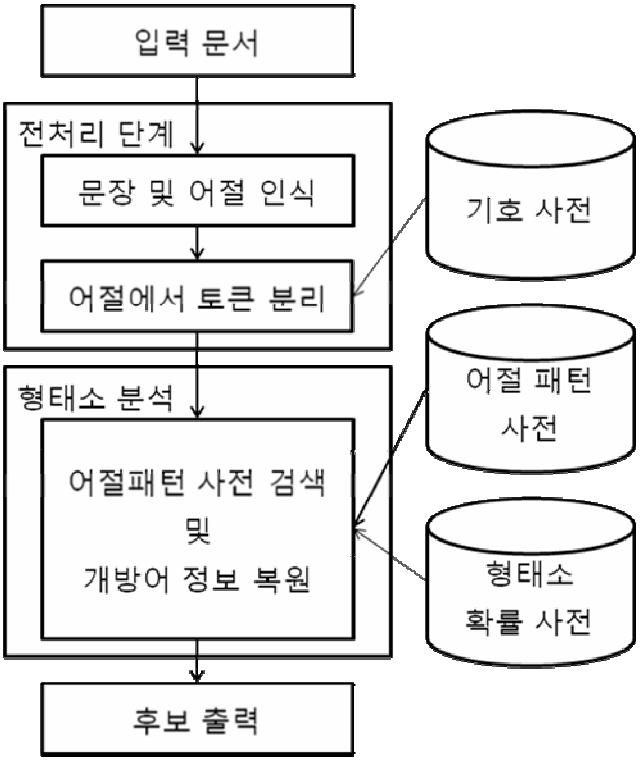

그림 1. KGuru-MA 처리 과정

3.1 전처리 단계

이 단계에서는 입력된 문서로부터 문장 단위로 분리하고 이를 다시 어절 단위로 분리한다. 여기서 어절에서 토큰(token) 단위로 분리를 하는데, 여기서 토큰의 정의는 그림 2와 같다.

1. 연속된 한글은 한 토큰이다.
2. 연속되는 숫자 문자열은 한 토큰이다.
3. 연속되는 알파벳 문자열과 알파벳 사이의 연결 문자(“.”, “-”, “_”, “|”)는 한 토큰이다.
4. 연속되는 한자는 한 토큰이다.
5. 하나의 기호는 한 토큰이다.
6. 그 외의 연속되는 외래 문자는 한 토큰이다.

그림 2. 토큰에 대한 정의

전처리 단계는 형태소 분석 과정에 들어감에 있어 먼저 처리할 입력 단위를 나누어 줌으로써 다음 단계의 처리를 용이하게 하는 데 그 의의가 있다. 특히, 연속된 한글로 구성된 토큰이 아니라면, 세종 말뭉치의 특성상 해당 토큰에 품사 정보를 바로 부여함으로써 형태소 분석을 보다 효율적으로 처리할 수 있다.

3.2 형태소 분석

형태소 분석 단계는 일반적인 형태소 분석기에서와 같이 형태소를 분석하는 단계로 특이한 점은 접속 검사 단계가 없다는 것이다.

이전 단계에서 말했듯이 한글로 구성되지 않은 토큰은 모두 해당 문자 토큰의 특성으로부터 품사 정보를 부여한다. 예를 들어, “love”, “あい”는 “외국어”, “愛”는 “한자”, “123”은 “수사”로 부여를 한다. 이 외 한글로 구성된 토큰의 경우에는 어절패턴 사전을 검색한다. 이 사전은 해당 패턴에 대해 개방어 정보를 “~”로 치환된 형태로 해당 어절패턴에 대한 모든 형태소 분석 정보를 담고 있다. 즉, “~는”의 대해서는 “~/명사+는/조사”,“~/고유명사+는/조사”, “~/동사+는/어미” 등의 정보를 가지고 있다는 것이다.

여기서 한글로만 구성된 토큰에 대해 분석될 수 있는 모든 형태소 정보에 대한 어절패턴을 찾아야 한다. 이 말은 이 과정 처리의 입력으로 전처리 결과값인 “상가가”와 같은 토큰이 입력으로 들어오는데, 이로부터 적절한 어절패턴을 찾아야 하는 문제이다. 이러한 문제를 해결하기 위해 한국어와 같은 조사나 어미가 발달한 언어의 특성을 살려 <그림 3>과 같은 의사 코드로 해당 어절패턴을 유추해내어 그 어절패턴이 가지고 있는 모든 형태소 정보를 가져온다.

```
w = 전처리 결과값으로 나온 한국어 token
wp = w가 전체문장에서 몇 번째 token인가?
buf = “”
i = w의 길이 - 1
f = False
retS = Null
while ( i == 0 )
{
  buf = str_replace(w, w[0:i-1], “~”)
  retS = search_EojeolPattern(buf)
  if ( buf != Null )
  {
    Inseart_Graph(w, wp, retS);
    f = True;
  }
  i = i - 1;
}

if ( f != True )
{
  buf = UnknownEojeolPattern_Process(w);
  Insert_Graph(w, wp, buf);
}
```

그림 3. 어절패턴을 이용한 형태소 정보 추출 의사 코드

이 의사코드가 의미하는 바는 입력으로 들어온 한글로만 구성된 토큰에 대해 문자열의 맨 마지막에서부터 한 음절씩 놓아두고 “~”로 치환하여 해당 어절패턴이 사전에 존재하는 지 여부를 검색하여 사전에 존재하면, 해당 어절패턴에 대한 형태소 분석 정보를 모두 그래프에 넣는다. 즉, “상가가”와 같은 토큰이 들어왔다면, 처음에 “상가가”에서 “상가”를 “~”로 치환한 형태인 “~가”에 대한 어절패턴을 찾고, 다음에는 “~가가”에 대한 정보를, 마지막으로 “상가가”에 대한 정보를 찾는다. 만약, 이렇게 유추한 어절패턴 후보에 대해 사전에서 정보가 존재하지 않는다면 해당 토큰 전체를 미등록 어절패턴 처리 모듈의 결과를 그래프에 삽입한다.

그러나 위와 같은 방법으로 형태소 후보를 찾게 되면, (1)패턴으로 인한 잘못된 형태의 후보 생성, (2)기존의 형태소 분석기에 비해 매우 많은 형태 분석 후보가 생성되는 문제가 있다. 이러한 문제를 해결하기 위해 본 연구에서는 확률 정보를 이용해 각 형태 분석 후보에 대해 순위를 매겨 최고 높은 n 개의 후보 만을 출력하도록 했다.

3.2 미등록 어절패턴의 처리

KGuru-MA에서는 미등록 어절패턴은 어절 전체를 명사 혹은 고유명사로 처리하는 방법을 취하고 있다.

이는 기존의 형태소 분석기에서 사용하는 “미등록어 처리 모듈”과는 상당히 차이가 있다. 그 이유는 3.2절에서 설명한 어절패턴으로 구성한 사전과 알고리즘으로 기존 형태소 분석기에서 처리하고 있는 미등록어 처리를 할 수 있기 때문이다. 어절 패턴에 의해 미등록어 처리를 해야하는 대부분의 경우는 알고리즘으로 처리되기 때문에 그 외의 결과는 어절 전체가 명사 혹은 고유명사라고 고려할 수 있다. 그래서 알고리즘에 의해 형태소 정보 추출이 실패할 경우, 이를 명사 혹은 고유명사로 처리하는 방법을 취했다.

4. 어절패턴 사전/형태소 확률 사전의 구축

어절패턴 사전과 형태소 확률 사전 구축은 본 형태소 분석기의 핵심 부분 중의 하나로 이미 구축된 품사 부착 말뭉치(Annotated Corpus)로부터 반자동 학습을 통해 구축 가능하다는 것이 가장 큰 특징이다. 이 과정의 전반적인 흐름은 그림 4와 같다.

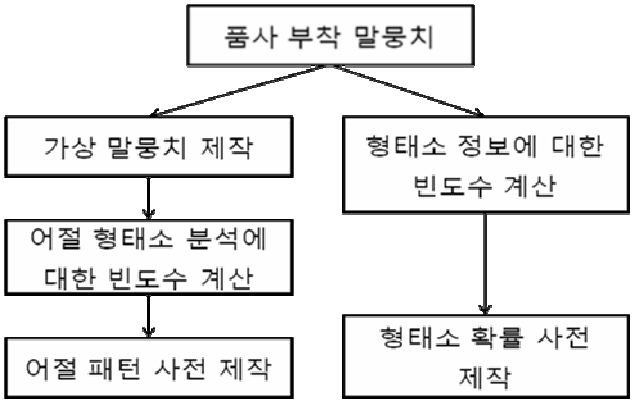

그림 4. 어절 패턴 및 확률 사전 구축 과정

형태소 확률 사전의 경우, 기존 확률 기반의 품사 태거에서 이용하는 확률 사전의 구축 방법과 같은 방법으로 품사 부착 말뭉치로부터 형태소에 대한 빈도수를 계산하여 확률 사전을 구축했다.

어절패턴 사전은 품사 부착 말뭉치로부터 가상 말뭉치로 변환한다. 가상 말뭉치란 그림 5의 가상 말뭉치의 예제와 같이 어절패턴에 맞게 변환한 말뭉치로 먼저 형태소 분석기에서의 전처리 과정과 같이 어절 및 토큰 분리 과정을 거치고 이로부터 한국어로 구성된 토큰에서 개방어의 형태소를 제거하되 활용형은 변화된 것을 모두 포함해 개방어가 없어진 자리는 "~"로 넣은 말뭉치를 의미한다. 예를 들어, "살렸다"라는 형태소 정보가 "살리/동사+었/선어말어미+다/종결어미"인 경우 어절 정보는 "~렸다"로 구성되고 형태소 정보는 "~리/동사+었/선어말어미+다/종결어미"와 같은 형식으로 변환된다. 이렇게 만들어진 가상 말뭉치로부터 어절이 형태소 분석이 되었을 때의 그 결과(이후, 이를 형태소 분석 패턴이라고 하며, 위의 예에서 "~리/동사+었/선어말어미+다/종결어미"를 의미)에 대한 빈도수를 측정하고, 어절패턴 단위로 형태소 분석 패턴을 그룹화 하여 그림 7과 같은 형식으로 구성한다. 이 때, 이 형태소 분석 패턴은 앞에서 형태소 분석 패턴에 대해 구한 빈도수가 높은 순 대로 나열한다.

그런데	그런데/MAJ
전세방	전세방/NNG
2	2/SN
개에	개/NNB + 에/JKB
8	8/SN
식구가	식구/NNG + 가/JKS
살고	살/VV + 고/EC
있더구먼	있/VX + 더구먼/EF
.	./SF
</s>	EOS/EOS
...	

품사 부착 말뭉치

그런데	그런데/MAJ
~	~/NNG
~	~/SN
개에	개/NNB+에/JKB
~	~/SN
~가	~/NNG+가/JKS
~고	~/VV+고/EC
있더구먼	있/VX+더구먼/EF
~	~/SF
~	EOS/EOS
...	

가상 말뭉치

그림 5. 말뭉치 예제

```
...
~가서      ~/NNG+가/VV+아서/EC|~/NNG+가/VV+서/EC
~가시려면 ~/VV+아/EC+가/VX+시/EP+려면/EC
~가아      ~/NNG+가/JKS+아/NA
~가야      ~/VV+아/EC+가/VX+아야/EC
~가에      ~/NNG+가/XSN+에/JKB|~/NNP+가/XSN+에/JKB|~/NNG+가/NNB+에/JKB|~/NNP+가/NNB+에/JKB
~가에는   ~/NNG+가/XSN+에/JKB+는/JX
~가에서   ~/NR+가/NNB+에서/JKB|~/NNG+가/XSN+에서/JKB
...
```

그림 6. 어절 패턴 사전

[형태소분석패턴1]|[형태소분석패턴2]|…|[형태소분석패턴n]

그림 7. 어절패턴 사전 데이터 구성 형식

5. 실험

본 논문에서는 실험에 "21세기 세종계획 3차년도 형태소 말뭉치"를 이용하였다. 이 말뭉치는 전체 문장수는 45만여 개이며, 어절수는 2백만여 개를 가지고 있다. 해당 말뭉치의 장르별 구성은 다음과 같다.

표 1. 세종 3차년도 말뭉치 장르별 구성

	소설류	비소설류	신문기사
어절수	512,652	1,059,715	447,956
문장수	48,038	73,843	21,340

실험 순서는 (1)형태소 분석기의 어절패턴 사전 및 확률 사전을 구축을 위한 학습용 말뭉치와 형태소 분석기 성능 분석을 위한 평가용 코퍼스를 구축하고, (2) 어절패턴 사전 및 확률 사전을 구축한다. 그리고 (3)형태소 분석 후보의 수를 n 개로 제한하여 해당 모델에 대한 형태소 분석기의 성능을 평가하는 순서로 진행했다.

먼저, 학습 및 평가용 말뭉치를 표 2와 같이 세종 말뭉치를 나눠 실험을 했다.

표 2. 학습 및 평가용 말뭉치 구성

	구성방법	어절수	문장수
학습용	(전체)-(평가용)	1,616,239	111,443
평가용	전체에서 임의 추출	404,084	31,778

이 학습용 말뭉치를 이용해 어절패턴 사전을 구축한 결과, 약 3만여 개의 어절패턴을 획득하였고 이에 대한

형태소 분석 패턴은 약 4만 1천여 개를 획득했다.

형태소 분석기의 성능 평가는 "태깅 정답 제시율"을 이용했다[13]. 이 방법은 형태소 분석기의 분석 결과 중 정답이 포함되어 있으면 맞도록 하여 일종의 제현률을 측정하는 방법으로 아래의 수식 (1)을 이용하여 계산했다.

$$\text{정답제시율} = \frac{\text{정답과\_일치되는\_분석이\_포함된\_토큰}}{\text{전체\_토큰의\_수}} \cdots (1)$$

성능 평가는 형태소 분석 후보의 개수를 임의의 n개에 대해 제한하는 방식으로 실험했다. 실험 방법은 후보의 수를 최대 12개에서 최소 5개까지 줄여보는 방법으로 실험했다. 이렇게 후보의 수를 5개에서 12개 내외로 한정한 이유는 기존의 형태소 분석기 시스템에서의 후보의 출력 결과로 미뤄볼 때, 이 정도의 후보의 수를 가져야 상위 자연언어 처리 시스템의 성능이 보장되기 때문이다. 그 결과, 후보의 수는 점차적으로 줄인 결과 성능은 미비하게 변화하였지만, 최대 12개를 후보를 가졌을 때의 성능과 최소 5개의 후보를 가졌을 때는 어느 정도 성능의 차이가 있음을 확인할 수 있었다. 표 3은 KGuru-MA의 성능 측정 결과이다.

표 3. 성능 평가 결과

	후보 5개	후보 12개
태깅정답제시율	96.990% (476314/491091)	97.666% (479632/491091)

6. 결론 및 향후 연구 방향

6.1 결론

본 연구에서는 품사 부착 말뭉치로부터 어절 패턴과 확률 정보를 자동으로 학습하여 이 정보를 이용한 형태소 분석기를 구현하였다.

제안한 KGuru-MA는 기존의 형태소 분석기에서 항상 수동으로 구축한 어휘 사전과 접속 정보 사전과 복잡한 한국어 특성에 대한 정보를 이용하지 않고 패턴을 이용해서도 형태소 분석기 구현이 가능하다는 점을 보여 주고 있다. 그리고 어절 패턴을 사용함으로써 기존의 형태소 분석기와 같이 시스템에 대한 복잡한 지식이 없는 사람이라고 하더라도 어절 패턴 추가를 위해 형태소 분석기의 성능을 향상 시킬 수 있다는 가능성을 제시했다. 또한, 이러한 단순한 방법과 적은 사전의 사용은 하위 자연언어 처리 시스템에서 요구하는 강건성을 제공하고 상위 자연언어 처리 시스템에서 해당 형태소 분석과 관련된 시스템의 부담을 크게 줄였다는 점에 그 의의를 둘 수 있다.

6.2 향후 연구 방향

본 연구에서 개발한 형태소 분석기는 확률 정보를 가지고 있는 형태소 분석기이다. 이 확률 정보는 현재로는 최적의 형태소 분석 후보를 찾기 위한 랭킹 정보에만 이용되고 있지만, 실제 이 사전은 품사 태거의 확률 정보 사전과 거의 동일한 구조로 방법으로 구성했다. 이 확률 정보를 현재의 형태소 분석기와 품사 태거의 확률 정보로 이용할 수 있도록 접목 시킨다면, 다음 단계의 자연언어 처리 시스템인 품사 태거 개발 시의 사전 추가의 부담이 없기 때문에 시스템 리소스 부담이 크게 덜어줄 수 있는 품사 태거의 개발이 가능할 것으로 기대된다.

참고 문헌

[1] K. Koskernnienmi, "Two-level Model for Morphological Analysis", Proceedings of IJCAI-83, pp.683-685, 1983.

[2] 강승식, 김영택, "한국어 형태소 분석기에서 선어말어미의 분석 모형", 정보과학회논문지, 제 18권, 제 5호, pp.505-513, 1991.

[3] 강승식, "한국어의 형태론적 특성과 형태소 분석기법", 정보과학회논문지, 제 12권, 제 8호, pp.47-59, 1994.

[4] Cahill, L. J. "Syllable-based Morphology", Proceedings of COLING-90, Vol.3, pp.48-53, 1990.

[5] D.B. Kim, S.J. Lee, K.S. Choi, G.C. Kim, "A Two-level Morphological Analysis of Korean", Proceeding of COLING-94, Vol. 1, pp.535-539, 1994.

[6] H.C. Kwon, L.Karttunen, "Incremental Construction of a Lexical Transducer for Korean", Proceedings of COLING-94), Vol. 2, pp.1262-1266, 1994.

[7] 최형석, 이주근, "자연어 어절 처리의 앨고리즘", 한국정보과학회 추계 학술발표회 논문집, 제 11권, 제 2호, 1984.

[8] 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기", 한국정보과학회 인공지능연구회 춘계 인공지능 학술발표회 논문집, pp.133-147, 1987.

[9] 김남철, 서영훈, "한국어 형태소 분석기 CBKMA와 색인어 추출기 CBKMA/IX", 제 11회 한글 및 한국어 정보처리 학술발표 논문집, pp.50-59, 1999.

[10] 권오욱, 정유진, 김미영, 류동원, 이문기, 이종혁,

"음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사 태거", 제 11회 한글 및 한국어 정보처리 학술발표 논문집, pp.76-86, 1999.
[11] 김덕봉, 최기선, 강재우, "한국어 형태소 처리와 사전 - 접속정보를 이용한 한글 철자 및 띄어쓰기 검사기- ", 어학연구, 26권, 1호, pp.87-113, 1990.
[12] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위논문, 1993.
[13] 이재성, 박재득, 차건회, 박세영, "형태소분석기 및 품사 태거 평가대회(MATEC99) 개요", 제 11회 한글 및 한국어 정보처리 학술발표 논문집, pp.13-22, 1999.