

# 어휘별 분류기를 이용한 한국어 품사 부착의 성능 향상

최원중<sup>0</sup>, 이도길<sup>\*\*</sup>, 임해창<sup>\*</sup>  
<sup>\*</sup> 고려대학교 컴퓨터학과 자연어처리연구소  
{wjchoi<sup>0</sup>, dglee, rim}@nlp.korea.ac.kr  
<sup>\*\*</sup>NLP Development Team, NHN corp.

## Improving Korean Part-of-Speech Tagging Using The Lexical Specific Classifier

Won-Jong Choi<sup>0</sup>, Do-Gil Lee<sup>\*\*</sup>, Hae-Chang Rim<sup>\*</sup>  
<sup>\*</sup>Natural Language Processing Lab., Korea University  
<sup>\*\*</sup>NLP Development Team, NHN corp.

### 요 약

한국어 형태소 분석 및 품사 부착을 위해 지금까지 다양한 모델들이 제안이 되었으며 어절단위 평가로 95%를 넘는 성능을 보여주는 자동 태거가 보고 되었다. 하지만 형태소 분석 및 품사 부착은 모든 자연어처리 시스템의 성능에 큰 영향을 미치므로 작은 오류도 중요하다. 본 연구에서는 대상 어절의 주변 형태소의 어휘와 품사 자질, 그리고 어절 자질을 이용하여 분류기를 학습한 후 자동 태거의 품사 부착 결과를 입력으로 받아 후처리 하는 어휘별 분류기를 제안한다.

실험 결과 어휘별 분류기를 이용한 후처리만으로 어절단위 평가 6.86%(95.251% → 95.577%)의 오류가 감소하는 성능향상을 얻었으며, 기존에 제안된 품사별 자질을 이용한 후처리 방법과 순차 결합할 경우 16.91%(95.251% → 96.054%)의 오류가 감소하는 성능 향상을 얻을 수 있었다. 특히 본 논문에서 제안하는 방법은 형태소 어휘까지 정정할 수 있기 때문에 품사별 자질을 이용한 후처리 방법의 성능을 더욱 향상시킬 수 있다.

### 1. 서론

한국어 형태소 분석 및 품사 부착은 그림 1과 같이 주어진 문장에 대하여 문장을 구성하는 각 어절에 대응하는 적절한 형태소, 품사열을 결정하는 것이다. 또한 한국어는 교착어(agglutinative language)로서 하나 이상의 형태소가 결합하여 어절을 이룰 수 있기 때문에 중의성이 매우 높다.

이러한 한국어 형태소 분석 및 품사 부착은 자연어를 이해하여 처리하는데 있어서 가장 기본적인 기술이어서 구문 분석, 기계번역, 정보 검색, 질의 응답 등 다른 모든 자연어처리 시스템에 사용된다. 따라서 품사 부착의 오류는 다른 상위 시스템에 전파되기 때문에 자연어처리 시스템의 성능에 영향을 미치는 부분이다.

지금까지 최고의 성능을 보이는 한국어 품사 태거의 경우 어절 단위 평가로 95% 이상의 성능을 보이고 있다[1]. 이러한 성능은 한국어의 교착어적인 특성상 상당히 높은 것이라고 할 수 있으나, 좀 더 성능 향상이 이루어진다면 자연어처리 시스템들의 성능도 향상할 수 있다. 또한 최근 “21세기 세종계획”의 결과물로 대량의 품사 부착 말뭉치의 구축[2]이 진행됨에 따라 학습에 필요한 충분한 양의 말뭉치 습득이 쉬워졌다.

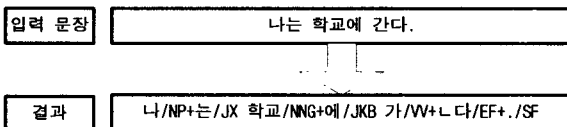


그림 1. 품사 부착의 예

이러한 한국어 품사 부착의 성능을 향상 시키기 위해서는 다양한 언어현상을 반영하는 자질들을 사용하여야 하나 기존에는 자료 부족 문제로 품사 문맥을 주로 사용하였다. 하지만 품사 문맥만으로는 성능 향상에 한계가 있다.

따라서 본 논문에서는 대량으로 구축된 말뭉치에서 얻어진 어절 어휘 문맥 자질과 기존 통계적 방식 태거의 결과에서 얻어진 형태소 어휘, 품사 문맥을 사용하여 품사 부착을 하는 어휘별 분류기를 제안한다.

제안된 어휘, 품사자질을 이용한 어휘별 분류기는 형태소 어휘와 품사 자질을 얻기 위하여 통계적 방식의 자동 태거[1]을 이용하나 추가적인 노력 없이 다른 태거와도 쉽게 결합하여 성능 향상을 얻을 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 품사 부착 후처리와 관련된 기존 연구를 기술한다. 3절에서는 한국어 품사 부착의 오류에 대해서 분석을 하며 4절에서는 제안하는 분류기에 대하여 설명한다. 5절에서는 분류기가 사용할 자질에 대하여 기술하며 6절에서는 제안한 분류기와 자질을 이용하여 향상되는 성능을 실험을 통하여 밝힌다. 마지막으로 7절에서는 결론 및 향후 연구 방향에 대하여 기술한다.

## 2. 기존 연구

품사 부착의 오류가 있는 부분을 효과적으로 찾고, 교정하여 성능을 향상하는 기존연구는 다음과 같은 것들이 있다.

[3]은 오류 가능성이 높은 부분을 자동으로 인식하기 위하여 두 가지 방법을 제안하였다. 첫번째는 널리 쓰이는 통계적 태거가 내어주는 N-Best의 결과 중에서 상위 품사 후보들간의 확률 차이가 근소할 경우 중의성이 높은 것이기 때문에 오류 가능성이 높다고 인식하는 방법이며, 두번째는 자동 태거의 결과를 분석하여 오류가 많이 발생하는 문맥을 정규표현식을 이용하여 규칙으로 구축해두고, 통계적 태거의 결과와 비교하여 구축된 규칙과 일치하는 부분을 오류 가능성이 높은 부분으로 인식하는 방법이다.

[4]는 통계적으로 오류가 많은 상위 10개의 어절을 대상으로 하여 좌우어절내의 대표 형태소 어휘 문맥 정보에 기반하여 오류를 정정하는 방식이다. 대표 형태소 어휘 문맥으로는 중심 어휘와 조사 어휘를 사용한다.

하지만, 대상 어절의 절반 이상의 정확률이 후처리 후에도 90%이하이며 중심 어휘와 조사 어휘 문맥이 중의성 해결에 도움이 되지 않는 어절에는 적용할 수 없으며 주변 어절 문맥을 반영할 수 없다는 단점이 있다.

[5]는 통계적 방식으로 품사 부착을 한 결과를 변형 기반 학습방법[6]을 이용하여 후처리 규칙을 학습한

다. 이때 학습된 규칙은 bigram HMM이 다루는 문맥보다 넓은 문맥을 다루기 때문에 HMM을 보완한다.

그러나 이 방식은 학습에 시간이 많이 걸리고, 특정 태거의 결과에 종속된다는 단점이 있다.

[7]은 중의성이 높아 오류 가능성이 높은 품사쌍을 대상으로 각 품사쌍에 적합한 자질을 이용하여 자동 태거의 품사 부착 결과에 후처리를 하여 품사 부착의 성능을 향상하는 품사별 자질을 이용한 후처리 방법을 제안하였다.

하지만 후처리를 통하여 품사만 정정하기 때문에 형태소 어휘가 틀린 경우는 정정할 수 없다는 한계가 있다. 또한 대상이 되는 중의성이 높은 품사쌍 이외에는 적용할 수 없다는 단점이 있다.

[8]은 정답이 있는 말뭉치에서 정답을 제거한 후 자동태거로 품사 부착하여 그 결과와 정답을 이용하여 통계적 결정 그래프 학습 방법으로 규칙을 학습하고, 그 규칙들을 후처리에 적용하는 방식이다.

이 방식은 규칙의 수가 많아질 경우 규칙 탐색에 시간이 많이 걸린다는 것과 학습된 규칙이 특정 태거의 후처리에만 이용될 수 있다는 단점이 있다.

이들 기존연구에 비해 제안하는 방법은 대상으로 하는 어절의 종류가 다양하며 후처리의 입력으로 사용하는 태거의 품사 부착 결과에 관계 없이 적용 가능하

## 3. 한국어 품사 부착 오류 분석

### 3.1 한국어 품사 부착 오류 유형

한국어는 여러 개의 형태소가 결합하여 하나의 어절을 구성할 수 있기 때문에 한국어 품사 부착에서 그림 2의 예에서와 같이 여러 개의 형태소 분석 후보가 가능하다.

하/VV+L/ETM  
 하/VX+L/ETM  
 한/MM  
 한/NA  
 한/NNG  
 한/NNP

그림 2. "한"의 형태소 분석 후보

따라서 품사 부착에서의 오류 유형은 그림 3과 같이 크게 두 가지로 나눌 수 있다.

유형 ①은 형태소 분리는 맞으나 품사가 틀린 경우로 품사만 정정을 하면 된다. 이 유형에 대한 후처리는 품사별 자질을 이용한 품사 부착 후처리[7]에서 연구가 되었다. 유형 ②는 형태소 분리가 잘못된 경우로 형태소 어휘가 틀렸으나 품사는 맞은 경우와 형태소

표 1. 한국어 품사 부착 오류 유형

오류 유형	오류 중 비율
유형 ①	44.30 %
유형 ②	55.70 %

표 2. 오류어절 중 학습집합 출현/미출현 비율

오류 유형	오류 중 비율
학습집합 출현 어절	43.3 %
학습집합 미출현 어절	56.7 %

- ① 형태소 분리는 맞으나 품사가 틀린 경우  
 예 : 하는 하/VX + 는/ETM [태거]  
 하/VV + 는/ETM [정답]
- ② 형태소 분리가 잘못된 경우  
 예 : 즐지 즐/VV + 지/EC [태거]  
 주/VV + 르지/EC [정답]

그림 3. 품사 부착 오류 유형

어휘와 품사가 모두 틀린 경우가 모두 이 유형에 속한다. 본 연구에 제안하는 어휘별 분류기는 유형 ①과 유형 ②를 모두 후처리 대상으로 삼는다.

### 3.2 한국어 품사 부착에서 오류 유형별 비중

세종계획 2003년도 품사 부착 말뭉치에 통계적 방식의 자동 태거[1]로 품사 부착을 하여 분석한 결과 각 오류 유형의 비율은 표 1과 같다.

표 1에 따르면 형태소 분리는 맞으나 품사만 틀리는 유형 ①이 45%, 형태소 분리가 잘못된 유형 ②의 경우가 55%로 형태소 분리가 잘못되는 경우가 더 많다. 또한 오류 어절 중 학습 집합에 출현한 어절과 그렇지 않은 어절의 비율을 비교하면 표 2와 같다.

즉, 오류의 상당 부분이 학습 집합에 출현 하였으나 높은 증의성으로 인하여 품사 부착시 잘못 선택된 것이다. 이러한 어절들 중 대부분은 어휘 문맥을 이용하면 정확하게 결정할 수 있으나 기존의 태거가 주로 주변 품사 문맥을 사용할 뿐 어휘 문맥과 좀더 넓은 범위의 품사 문맥을 사용하지 않기 때문이다. 따라서 기존의 태거가 이용하지 않는 자질인 형태소 어휘 문맥, 품사 문맥, 어절 문맥을 고려한다면 좀 더 정확하게 품사 부착을 할 수 있는 가능성이 있다.

## 4. 어휘별 분류기

어휘별 분류기를 이용한 품사 부착 후처리는 주어진 어절에 대하여 주변 형태소의 어휘와 품사 그리고 어

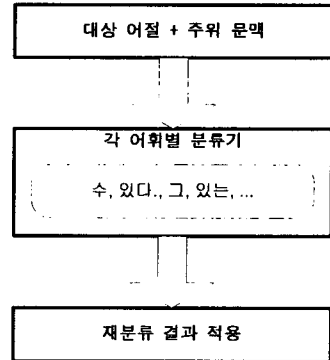


그림 4. 어휘별 분류기를 이용한 품사 부착 후처리

절 자질을 이용하여 새로운 품사 부착 결과를 결정하는 것이다. 이때 후처리 대상은 학습 말뭉치에서 일정 빈도 이상으로 출현한 어절을 대상으로 새로운 부착 결과의 후보는 학습 집합에 출현한 품사 부착 결과를 대상으로 한다. 출현빈도가 낮은 어절을 제외하는 것은 충분한 양의 학습 예제를 얻기 위함이다.

즉 후처리 문제는 학습 말뭉치에 출현한 후보들 중에서 주어진 어절과 형태소 자질을 기반으로 가장 적합한 것을 선택하는 분류 문제라고 할 수 있다.

따라서 본 논문에서는 그림 4와 같이 다른 태거의 품사 부착 결과를 입력으로 받아서 대상 어절과 주위 문맥을 추출한 후, 대상 어절의 어휘별로 학습된 분류기를 이용하여 새로운 품사를 결정하여 반영한다.

이때 입력으로 사용 가능한 태거는 규칙기반, 통계기반 등 모든 태거의 입력을 사용할 수 있으며, 각 어휘별 분류기를 학습하기 위해서는 자연어처리 분야에서 높은 성능을 보이며 여러 자질을 결합하기 용이한 최대 엔트로피 모델(Maximum Entropy Model)을 이용한다.

## 5. 자질

어휘별 분류기는 어휘의 종류가 다양하고, 각 어휘에 적합한 자질을 일일이 밝히기 어렵기 때문에 각 어휘에 적합한 자질을 이용하기 어렵다. 따라서 어휘별 분류기에서는 사용 가능한 모든 자질을 이용하는데, 사용하는 자질은 다음과 같다.

- ① 형태소 자질
  - 형태소의 어휘
  - 형태소의 품사
- ② 어절 자질
  - 어절의 어휘

이때 사용되는 문맥의 범위는 형태소 자질의 경우 대

표 3. 상위 20개의 고빈도 오류 어절

어절	정확률
한	72.488 %
한다.	86.047 %
하는	87.149 %
내	87.387 %
했다.	87.919 %
할	89.855 %
하고	90.278 %
볼	90.909 %
다른	91.903 %
안	93.529 %
이	94.459 %
그런	95.436 %
이를	95.489 %
수가	95.833 %
따라서	96.032 %
일이	96.403 %
보면	96.947 %
있을	97.126 %
등을	97.484 %

표 4. "한", "다른"의 정보 이득(Information Gain)

자질	"한"	"다른"
형태소 품사(-4)	0.136	0.031
형태소 품사(-3)	0.233	0.042
형태소 품사(-2)	0.373	0.109
형태소 품사(-1)	0.499	0.097
형태소 품사(+1)	0.145	0.020
형태소 품사(+2)	0.118	0.037
형태소 품사(+3)	0.089	0.043
형태소 품사(+4)	0.075	0.029
형태소 어휘(-4)	0.647	0.174
형태소 어휘(-3)	0.638	0.214
형태소 어휘(-2)	0.827	0.249
형태소 어휘(-1)	0.523	0.161
형태소 어휘(+1)	1.018	0.249
형태소 어휘(+2)	0.486	0.098
형태소 어휘(+3)	0.646	0.222
형태소 어휘(+4)	0.494	0.150
어절 (-3)	0.958	0.294
어절 (-2)	1.105	0.320
어절 (-1)	1.226	0.335
어절 (+1)	1.157	0.326
어절 (+2)	1.064	0.337
어절 (+3)	0.931	0.277

상 어절의 좌우 각 4개의 형태소를, 어절 자질의 경우 형태소 자질에 비해서 데이터 부족 문제가 심하기 때문에 대상 어절의 좌우 각 3개의 어절을 이용한다.

### 5.1 형태소 자질

다른 태거의 품사 부착 결과를 형태소 자질로 이용하기 위해서는 그 결과가 신뢰할 만한 수준이어야 한다. 기존의 태거들의 성능이 어절 단위 평가로 이미 95% 이상을 상회하고 있으며 형태소 단위 평가를 하면 97%를 상회하고 있다. 따라서 기존 태거의 품사 부착 결과에서 추출한 형태소 어휘 자질 및 품사 자질을 이용할 수 있다.

### 5.2 각 자질들의 분류 능력

각각의 자질이 분류에 어느 정도 도움이 되는지 평가하기 위하여 평가 척도로 정보 이득(Information Gain)을 이용하며, 정보 이득은 다음과 같이 정의한다.

$$\begin{aligned}
 I(Y|X) &= H(X) - H(Y|X) \\
 &= \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}
 \end{aligned}$$

즉  $H(X)$ 는 확률변수  $X$ 의 불확실성이며,  $I(Y|X)$ 는  $Y$ 를 알게 된 후의  $X$ 의 감소된 불확실성이다. 이 정보 이득을 이용하여 자질들의 분류능력을 알아보기 위하여 오류가 많은 어절들의 정보 이득을 구해보았다. 오류가 많은 어절은 통계적 방식의 자동태거[1]의 품사 부착 결과 중 출현빈도가 100이상인 어절에서 오류 상위 20개를 선택하였으며 그 어절들은 표 3과 같다.

표 3에서 나타난 고빈도 오류 어절 중에서 "한"과 "다른"의 정보 이득을 구하면 표 4와 같으며 다른 어절들도 같은 경향을 보인다. 수치가 높을수록 분류에 더 도움이 되는 자질이다.

표 4에서 형태소 어휘 혹은 형태소 품사는 형태소 자질이며 어절은 어절의 어휘 자질이다. 괄호 안의 숫자는 대상 어절로부터의 위치를 나타내며 -는 왼쪽, +는 오른쪽 문맥이다. 표 4에 따르면 어휘 자질이 품사 자질 보다, 어절 어휘 자질이 형태소 어휘 자질 보다 더 분류에 도움이 된다.

또한 어절의 어휘 자질의 경우 대상 어절에 인접할수록 분류에 더 도움이 되었으나, 형태소 어휘 자질의 경우는 주변에 인접한 정도와 상관없이 분류에 도움이 되는 정도가 달랐다. 하지만 형태소 품사 자질의 경우는 주변에 인접할수록 분류에 더 도움이 되었다.

표 5. 어휘별 후처리 분류기의 정확률

어절	출현 빈도	자동태거의 정확률	후처리 분류기의 정확률
한	647	72.488 %	88.099 %
한다.	559	86.047 %	90.877 %
하는	498	87.149 %	88.353 %
내	222	87.387 %	94.144 %
했다.	298	87.919 %	93.960 %
할	552	89.855 %	96.377 %
하고	288	90.278 %	93.056 %
볼	154	90.909 %	97.403 %
다른	247	91.903 %	97.166 %
안	170	93.529 %	98.824 %

## 6. 실험

총 200만 어절로 구성된 세종계획 2003년도 품사 부착 말뭉치 데이터를 사용하였으며, 문장당 평균 어절 수는 14.4 어절이다. 학습에는 90%인 180만개의 어절을, 실험에는 10%인 20만개의 어절을 사용하였다.

최대 엔트로피 확률을 추정하기 위해서는 Zhang Le의 Maximum Entropy Modeling Toolkit[9]을 이용하여 학습을 수행하였다. 파라미터 추정은 L-BFGS 알고리즘을 사용하고, 반복회수는 150으로 하였다.

### 6.1 학습된 분류기의 성능

출현빈도 100이상인 어절들 중에서 오류율 상위 10개의 어절을 대상으로 어휘별 분류기가 기존의 통계적 방식의 자동 태거보다 정확하게 분류를 하는지 실험하였다. 평가는 다음과 같은 어절단위 평가를 하였다.

$$\text{어휘별 정확률} = \frac{\text{정답과 일치하는 어절의 수}}{\text{어절의 출현 빈도}}$$

표 5를 보면 학습된 어휘별 분류기가 통계적 방식의 기존 태거에 비하여 더욱 높은 정확률로 분류함을 알 수 있다. 어절 “한”의 경우 통계적 방식의 기존 태거가 72.488%의 정확률인 반면 학습된 어휘별 분류기의 경우 88.099%로 56.74%의 오류가 감소하였다. 또한 어절 “내”의 경우는 87.387%에서 94.144%로 49.57%의 오류가 감소하였다. 이는 학습된 어휘별 분류기가 통계적 방식의 기존 태거보다 높은 정확률로 품사 부착을 한다는 것을 의미하며 어절 어휘의 출현 빈도에 따라 전체 성능에 미치는 영향의 정도는 차이가 있을 수 있지만, 결국 주위 문맥 자질을 이용한 어휘별 분류기를 적용하면 한국어 품사 부착의 성능을 향상할

표 6. 후처리 적용 후 성능 비교

방식	정확률
1. 통계적 방식의 자동태거 [1]의 결과	95.251%
2. 1의 결과에 어휘별 분류기 후처리 적용 결과	95.577% (+ 0.326%)

수 있다는 것을 의미한다.

### 6.2 후처리에 적용

높은 정확률을 가지는 어휘별 분류기를 적용할 경우를 실험 하였다. 입력으로는 통계적 방식의 자동태거 [1]의 결과를 이용하였으며, 학습집합 출현빈도 100이상인 어절을 대상으로 후처리를 적용하였다. 이때 적용되는 어절의 비율은 실험 데이터의 34.2%이고, 성능 평가는 다음과 같이 어절 단위 평가를 사용하였으며, 후처리 결과의 성능은 표 6과 같다.

$$\text{정확률} = \frac{\text{정답과 일치하는 어절의 수}}{\text{전체 어절의 수}}$$

학습된 어휘별 분류기를 적용한 결과 어절 단위 성능 평가로 0.326%의 성능 향상으로 6.86%의 오류가 감소하였다. 이는 어휘 자질의 데이터 부족 문제를 감안할 때 의미 있는 성능 향상이라고 할 수 있다.

기존의 후처리 방법[7]이 표 1의 오류 유형 중 유형 ①만 후처리 대상으로 할 수 있는데 비하여 어휘별 분류기는 유형 ①과 유형 ②의 오류를 모두 후처리 대상으로 할 수 있다.

또한 기존의 후처리 방법[7]이 형태소의 품사만 지정하는 반면에 어휘별 분류기는 형태소의 품사뿐만 아니라 형태소의 어휘도 지정할 수 있다. 따라서 어휘별 분류기로 더욱 정확하게 후처리된 형태소 어휘 자질은 기존의 후처리 방법과 결합하여 성능을 더 향상할 수 있다.

### 6.3 품사별 자질을 이용한 품사 부착 후처리와의 순차 결합

본 연구에서 제안한 어휘별 자질을 이용한 품사 부착 후처리 방식과 기존에 제안된 품사별 자질을 이용한 품사 부착 후처리[7]를 순차 결합한 성능을 실험 하였다.

#### 6.3.1 품사별 자질을 이용한 후처리 우선 결합

품사별 자질을 이용한 후처리 우선 결합은 그림 5와 같이 통계적 방식의 자동 태거[1]의 결과에 품사별 자

표 7. 품사별 자질을 이용한 후처리 우선 결합

결합 순서	정확률
1. 통계적 방식의 자동태거[1]의 결과	95.251%
2. 1의 결과에 품사별 자질을 이용한 후처리 결과	95.872% (+0.621%)
3. 2의 결과에 어휘별 분류기를 적용한 결과	95.945% (+0.694%)

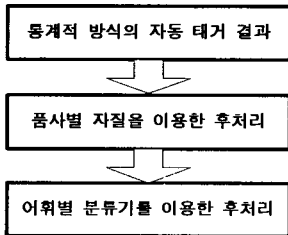


그림 5. 품사별 자질을 이용한 후처리 우선 결합

질을 이용한 후처리[7]를 적용한 후 어휘별 자질을 이용한 후처리를 적용한 결과이다. 성능은 어절단위 평가하였다.

실험 결과 95.251%에서 95.945%로 0.694%의 성능 향상이 있었으며, 14.61%의 오류가 감소하였다. 그러나 품사별 자질을 이용한 후처리 결과에 어휘별 분류기를 적용한 결과는 성능 향상이 크지 않았다. 그 이유는 품사별 자질을 이용한 후처리에서 정정하는 것은 품사 뿐이기 때문에 어휘 자질을 많이 사용하는 어휘별 분류기에 자동 태거[1]의 결과와 후처리 결과[7]가 큰 차이가 없기 때문이다.

### 6.3.2 어휘별 분류기를 이용한 후처리 우선 결합

어휘별 자질을 이용한 후처리 우선 결합은 그림 6과 같이 통계적 방식의 자동 태거[1]의 결과에 어휘별 자질을 이용한 후처리를 적용한 후 품사별 자질을 이용한 후처리[7]를 적용한 결과이다. 성능은 어절단위 평가하였다.

실험 결과 95.251%에서 96.054%로 0.804%의 성능 향상이 있었으며, 16.91%의 오류가 감소하였다. 어휘별 분류기를 이용한 후처리는 형태소의 품사와 어휘를 모두 정정할 수 있기 때문에 어휘별 분류기를 적용하면 형태소 어휘의 정확률이 향상된다. 따라서 어휘 자질을 많이 쓰는 품사별 자질을 이용한 후처리의 성능에도 도움이 된다. 그 결과 어휘별 분류기를 이용한 후처리 우선 결합이 품사별 자질을 이용한 후처리 우선 결합보다 성능이 높게 된다.

표 8. 어휘별 분류기를 이용한 후처리 우선 결합

결합 순서	정확률
1. 통계적 방식의 자동태거[1]의 결과	95.251%
2. 1의 결과에 어휘별 분류기를 적용한 결과	95.577% (+0.326%)
3. 2의 결과에 품사별 자질을 이용한 후처리 결과	96.054% (+0.803%)

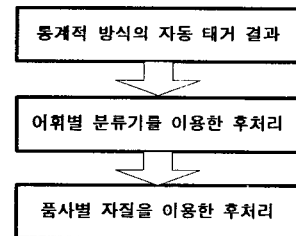


그림 6. 어휘별 분류기를 이용한 후처리 우선 결합

## 7. 결론 및 향후 연구

품사 부착의 성능을 향상하고자 할 때 자료 부족 문제가 심하지만 가장 유용한 자질은 어휘 자질이다. 또한 후처리 방식을 취하면 형태소 어휘 및 품사 자질을 이용할 수 있다.

본 연구에서는 어절 어휘 자질과 기존의 자동 태거에서 얻을 수 있는 형태소 어휘, 품사 자질을 품사 부착 후처리에 이용하여 0.326%의 어절단위 정확률 향상을 얻었다.

또한 본 연구에서 제안한 어휘별 분류기는 형태소의 품사뿐만 아니라 어휘도 정정할 수 있기 때문에 기존에 제안된 품사별 자질을 이용한 품사 부착 후처리 방법과 순차 결합하여 0.803%의 어절단위 정확률 향상을 얻었다.

학습 집합의 크기는 어휘 자질을 많이 사용하는 분류기의 특성상 크기가 클수록 성능 향상이 크겠지만, 180만 어절의 학습 말뭉치로도 의미 있는 성능 향상을 얻을 수 있었기 때문에 추후 더 큰 크기의 학습 말뭉치가 주어진다면 더욱 높은 성능 향상을 얻을 수 있다.

연구 결과 어휘별 분류기는 품사 부착 성능 향상에 도움이 되었으며, 추가적으로 전체 오류 중 절반 가량을 차지하며 현재 후처리 대상으로 하고 있지 않은 미등록 어절까지 처리할 수 있도록 하는 방법이 연구가 된다면 더욱 큰 성능 향상을 얻을 수 있을 것이다.

## 8. Acknowledgements

이 논문은 2005년 정부(교육인적자원부)의 재원으로  
한국학술진흥재단의 지원을 받아 수행된 연구임  
(KRF-2005-041-D00737)

## 9. 참고 문헌

- [1] 이도길, 한국어 형태소 분석과 품사 부착을 위한 확률 모형, 고려대학교 박사학위논문, 2005.
- [2] 강범모, “언어, 컴퓨터, 코퍼스 언어학: 컴퓨터를 이용한 국어 분석의 기초와 이론”, 고려대학교 출판부, 2003.
- [3] Elliott Macklovitch, "Where the Tagger Falsters", In Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation, pp 113-126, 1992.
- [4] 김영길, 양성일, 홍문표, 박상규, "형태소 어휘 문맥에 기반한 태깅 오류 수정", 제 15회 한글 및 한국어 정보처리 학술대회, 2003.
- [5] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2 단계 한국어 품사 태깅 시스템", 정보과학회논문지 제 24권 제 2호, 1997.
- [6] Eric Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case study in Part-of-Speech Tagging", Computational Linguistics, vol. 21, no. 4, pp. 543-564. 1995.
- [7] 최원중, 이도길, 임해창, "품사별 자질을 이용한 한국어 품사 부착의 성능 향상", 한국 컴퓨터 종합 학술대회 논문집 제 33권 제 1호, pp 16-18, 2006.
- [8] 류원호, 이상주, 임해창, "통계적 결정 그래프 학습 방법을 이용한 한국어 품사 부착 오류 수정", 제 13회 한글 및 한국어 정보처리 학술대회, 2001.
- [9] Zhang Le, "Maximum entropy modeling toolkit for Python and C++", 2004.