# 형태소 분석기를 위한 효율적인 미등록 명사 추정 알고리즘

신준철○, 옥철영
울산대학교, 울산대학교
ducksjc@nate.com, okcy@ulsan.ac.kr

# An Efficient Recognition Algorithm of the Korean Unknow-words for Morpheme Analyser

Joon-Choul Shin○, Cheol-Young Ock
Ulsan University, Ulsan University

**요 약**

한국어 자료를 자동으로 처리하기 위해서 다양한 형태소 분석기가 연구되었으나, 대부분의 형태소 분석기는 미리 등록된 명사가 아니면 제대로 분석하지 못하는 문제점을 가지고 있다. 본 논문은 기존의 형태소 분석기를 수정하여 미등록 명사를 인식하도록 하는 방법을 소개한다. 이 방법은 비록 학습 알고리즘을 포함하지 않지만 비교적 구현이 쉽고 속도가 빠르며 형태소 분석기의 정확률 향상에 도움이 되었음을 실험으로 검증하였다. 그리고 이 알고리즘을 응용하여 사람이 반자동으로 미등록 명사를 포함할 가능성이 높은 어절을 수집하는 방법을 제안한다.

주제어: 형태소 분석기, 형태소, 자연어처리, 한국어처리, 미등록어, 신조어

## 1. 서론

일반적으로 한국어 처리를 위해서는 형태소 분석기가 필요하기 때문에 다양한 형태소 분석기가 연구되고 있다. 한국어는 하나의 어절에 여러 형태소가 조합될 수 있는 교착어이기 때문에 자동 형태소 분석에 많은 어려움이 있다. 대표적으로 어려운 점으로는 조사와 명사의 경계를 구분하는 것과, 복합명사와 미등록 명사를 구분하여 처리하는 것이다. 이 문제들은 모두 동시에 발생할 수 있기 때문에 하나의 문제로 인식할 수도 있다.

대체로 형태소 분석기들은 미리 형태소들을 등록하고 이 정보를 적극적으로 이용하여 형태소 분석을 한다. 이런 특징 때문에 만약에 등록되지 않은 형태소를 만나게 되면 정확률이 급격히 낮아지게 된다. 대표적으로 발생하는 오류는 1음절 또는 2~3음절의 명사들로 구성된 복합명사로 인식하는 것이며, 가끔 조사와 명사의 경계를 구분하지 못하기도 한다. 이런 문제를 해결하기 위해서는 미등록 명사임을 추측하는 기술이 필요하다.

기존에도 미등록 명사 문제를 해결하기 위한 다양한 연구가 있으나 미등록 추측을 위한 추가적인 학습 정보를 필요로 한다. 본 논문은 최근에 연구된 형태소 분석기의 내부 학습사전만을 이용하는 방법을 소개한다.

## 2. 관련 연구

영어는 미등록어를 분석하더라도 하나의 단어가 하나의 품사와 형태소로 구성되므로 단어 내에서 미등록어의 경계를 인식할 필요가 없고 품사만 추정한다[1]. 이 때 품사 추정은 주로 문맥 정보와 해당 단어의 접두사, 접미사 그리고 대문자로 시작하는 정보 등을 이용한다. 국내의 영어 미등록어 추정에 관한 연구로는 김형철(2009)의 연구가 있다[2].

반면에 한국어는 하나의 어절이 여러 형태소로 구성될 수 있는 교착어이기 때문에 미등록 명사가 발생할 경우에 형태소 분석이 매우 어렵다. 우선 품사 추정을 하기 이전에 각 형태소의 경계를 구분해야하는 것이 순서상 옳바르지만 형태소의 경계 문제는 각 형태소의 품사 추정 문제와도 연결되기 때문에 주의를 요한다. 가장 일반적인 방법은 미등록 어절을 명사와 조사의 결합으로 보고, 어절의 오른쪽에서 최장 조사를 떼어내고 남은 부분을 미등록 명사로 인식하는 최장조사 분리 방법이다[3]. 김선호(2002)는 각 문서를 대상으로 형태소 분석 단계 이전에 미등록 형태소의 대상이 되는 단어들을 동적 사전으로 만들어 형태소 분석시에 사용하였다[4]. 우선 suffix array 구조를 이용해 문서로부터 어절들의 최장 공통 문자열을 추출하여 로컬 사전을 생성하고, 이 사전은 기존 형태소 사전을 보조하여 미등록어 분석에 도움을 준다. 실험 결과 이 방법은 미등록어 발생시 추정에 의한 과분석을 방지하고 보다 정확한 분석이 가능하도록 하였다.

위의 연구들과는 다르게 기계학습 방식을 사용한 연구도 있다. 박재한(2004)은 일반적인 복합명사와 미등록 외래어를 포함한 복합명사를 잘 분해하기 위해서 1,000만 어절의 세종말뭉치에서 448만개 명사와 복합명사를 분리해 놓은 것에서 백오프 통계 정보를 학습하여 사용하였다[5]. 통계 정보는 음절 바이그램, 어휘 바이그램,

- 본 연구는 2012년 2014년 정부(교육과학기술부)의 재원으로 한국연구재단 연구사업의 지원을 받아 수행된 연구임(No. 2012R1A1A2006906, 2014R1A1A2009506)

품사 바이그램 등으로 구성된다. 실험 결과 미등록 명사 추정 모듈에 의해 전체 시스템의 성능이 향상되는 것을 확인하였다. 최맹식(2011)은 SVM(Support Vector Machine)을 이용하여 미등록 형태소 오류들을 포함할 가능성이 있는 어절들을 검출하고 CRFs(Conditional Random Fields)를 이용하여 검출된 형태소 분리와 품사 태깅을 수행하는 방법을 제안하였다[1]. 제안된 모델을 실험하기 위해 세종 말뭉치 중 임의로(randome)하게 추출(sampling)한 139,757문장을 어절 단위로 실험하였고, 미등록 명사를 처리하지 않았을 때의 정확률 94.86%에서 0.35% 향상된 95.21%를 보였다. 사용된 형태소 분석기의 기본 정확률이 낮았으나 정확률 향상 정도가 높아 의미 있는 연구이다. 이런 기계학습 알고리즘은 비교적 정확률이 높지만 학습 과정을 필요로 하고 그 결과물인 학습 정보를 사용해야 작동이 가능하다. 따라서 기존 형태소 분석기와 함께 작동하기 위해서는 학습 시간과 메모리 사용 문제에 부담이 될 수 있을 것이다.

이렇게 대부분의 연구들은 미등록 명사를 포함하는 해당 어절만을 분석하여 미등록 명사를 추정하였지만 문맥 정보를 이용하여 미등록 명사를 추정하는 연구도 있다. 박봉래(1998)는 주어진 문서에서 동일한 미등록 명사가 사용된 용례를 찾아서 구의 비교를 통해 미등록 명사를 인식하는 방법을 제안하였다[6]. 이 방법은 문맥 정보를 이용하기 때문에 해당 미등록 명사가 교유명사인지 판정하는 기능도 포함한다. 이와 같은 특징들 때문에 이 연구는 다른 연구들에 비하여 매우 의미가 크다.

## 3. 미등록 명사 추정

### 3.1 기존 형태소 분석기

미등록 명사 추정 알고리즘은 기존 형태소 분석기와 함께 사용될 것이기 때문에 새로 설계할 때에는 그 알고리즘이 적용될 형태소 분석기를 파악하고 어울리도록 설계하는 것이 중요하다.

본 논문은 신준철(2012)의 기분석 부분 어절 사전을 활용한 분석기에 신준철(2014)의 부분어절 조건부 확률 기반의 태깅 모델을 조합한 UTagger를 사용한다[7, 8]. 이 형태소 분석기는 우선 1,100만 어절의 세종말뭉치를 학습하여 기분석 사전을 구축하고 어절을 분석할 때에 기분석 사전을 활용한다. 여기서 사용하는 기분석 사전은 말뭉치에 나타난 어절의 분석된 형태를 저장하고 있으며, 어절의 부분도 그에 해당하는 분석 정보와 함께 저장한다. 예를 들어 '사과를'의 기분석 내용은 <표 1>과 같다.

<표 1> '사과를'의 기분석 내용

| 분석 내용 | 빈도 | 형태 |
|---|---|---|
| 사과__05[apple]/NNG+를/JKO | 90 | 전체 |
| 사과__08[apology]/NNG+를/JKO | 88 | 전체 |
| 사과__05[apple]/NNG+를/JKO | 3 | 부분 |
| 사과__08[apology]/NNG+를/JKO | 2 | 부분 |

<표 1>에서 '사과를'이 어떤 어절의 부분으로 나타난 경우는 총 5번으로 3과 2를 더한 것이다. 여기서 '사과'의 의미가 apple인 것은 총 93개로 90과 3을 더한 것이다. <표 2>는 '를'의 기분석 내용이다. '를'은 어절의 전체로 등장한 것 보다 부분으로 등장한 경우가 많은 것을 확인할 수 있다.

UTagger는 만약 분석 대상 어절 전체가 기분석 사전에 등록되어 있지 않다면 어절을 두 개로 나누어서 기분석 사전을 검색하고, 두 개로 나누어서도 해결하지 못했다면 복합명사의 가능성이 있는 것으로 간주하여 더 잘게 나누어 분석을 시도한다.

어절을 두 개 이상으로 나눌 때 일반적으로 어절의 최우측 부분은 조사를 포함하는 형식 형태소이며 기분석 사전에서 발견하기 쉽다. 예를 들어서 '고급정보라면'은 세종말뭉치에 전체 어절이 등록되어 있지는 않지만

<표 2> '를'의 기분석 내용

| 분석 내용 | 빈도 | 형태 |
|---|---|---|
| 를/JKO | 275,963 | 부분 |
| 를/JKO | 11,729 | 전체 |

'~라면'은 다수 발견할 수 있다. 만약 미등록 명사가 포함된 어절이라면 최우측은 기분석 사전에서 발견하고 나머지는 발견하지 못했거나 발견하였더라도 매우 어색하거나 문법적으로 틀린 분석일 확률이 높다.

### 3.2 분석 후보의 생성과 점수

UTagger는 각 후보들을 생성할 때 문맥 정보를 고려하지 않고 오직 해당 어절 정보만을 고려하여 후보 점수를 계산한다. 이 점수는 비록 문맥 정보를 고려한 것은 아니지만 어절 내부의 자연스러움과 문법적인 가능성을 계산한 것이기 때문에 지나치게 낮다면 해당 후보는 태깅 단계로 넘어가기 전에 삭제될 수 있으며, 태깅 단계에서는 문맥 정보와 함께 사용되기 때문에 매우 중요하다[8]. 만약 분석 대상 어절의 모든 후보의 점수가 매우 낮다면 그 어절은 미등록 명사를 포함하고 있다고 추정할 수 있다. 따라서 특정 점수 이하로만 후보가 생성되었다는 조건을 통해서 미등록 명사 추정의 시도 여부를 판단한다.

UTagger에서 미등록 명사 추정을 한다는 것은 새로운 후보를 생성한다는 의미이며, 새로운 후보는 다른 후보와 마찬가지로 점수가 필요하다. 이 점수는 다른 후보들의 점수와 비교할 수 있도록 비슷한 방법으로 계산되어야 한다. 그렇게 된다면 미등록 명사 추정으로 생성된 후보의 점수와, 복합명사 추정으로 생성된 후보의 점수를 비교하여 해당 어절이 미등록 명사를 포함한 것인지 여부를 판단할 수도 있다.

일단 미등록 명사 추정을 시도하게 되면 어절의 최우측 부분의 기분석 정보가 조사인지 확인하고 활용하는 것이 우선이다. 이것은 최장조사 분리방법과 흡사하지만 반드시 최장일치를 사용하는 것은 아니다. 본 논문이 소개하는 방법은 기분석 사전에서 발견한 빈도 정보와 길이 정보를 같이 사용하여 "미등록 명사 추정 후보"의 점수를 계산하는 것이다.

UTagger는 일반적으로 어절의 좌측 분석의 빈도와 우측 분석의 빈도 그리고 그 둘의 조합 가능성 정도를 모

두 곱하여 후보 점수를 계산한다. 조합 가능성 정도란, 예를 들어서 명사 다음에는 조사가 올 확률이 높다는 것을 적용하기 위한 품사 바이그램 학습 정보와, 명사 끝 글자의 종성이 존재하면 조사 '을' 이 다음에 올 수 없다는 등의 규칙 정보들을 사용하여 계산된 수치다.

복합명사 추정 후보는 좌측 분석이 말뭉치에 나타난 적이 없는 새로운 복합명사로 추정된 것이기 때문에 빈도가 0이어서 일반적인 후보 점수 계산 방법을 적용할 수 없다. 따라서 좌측 빈도를 대체할 좌측 점수를 계산하여서 이것을 빈도처럼 이용한다. 빈도는 0또는 양의 정수이지만 추정된 복합명사의 점수는 대체로 0에서 1사이로 계산된다. 이 좌측 점수를 사용하여 후보 점수를 계산할 수 있다. 미등록 명사 추정 후보의 점수 계산에서도 미등록 명사로 추정되는 좌측 부분의 점수만 계산하면 UTagger의 기분석 사전 정보와 조합 가능성 정도를 통해 후보 점수를 계산할 수 있다.

결국 UTagger에서 미등록 추정 기능을 구현하기 위해서 가장 중요한 것은 어절 내에서 미등록 명사로 추정되는 좌측 부분에 대한 점수를 계산하는 것이다. 본 논문에서는 추가적인 학습과정 없이 간단하게 처리하는 방법을 소개한다. 조사부분의 길이와 미등록 명사 부분의 길이 정보만을 사용하여 미등록 명사 부분의 점수를 계산하는 것이다.

조사최장 방법과 유사한 가정을 하여서 조사부분의 길이가 길수록 점수가 높아지고, 반대로 미등록 명사 부분의 길이가 길수록 점수가 낮아지게 계산한다. 따라서 미등록 명사 부분의 점수 계산식은 식 (1)과 같다. $l$은 미등록 명사 부분 점수이며, $s_r$은 우측부분인 조사부분의 길이, $s_l$은 좌측부분인 미등록 명사 부분의 길이, 그 외에 a부터 e까지는 상수이다. 각 상수들은 해당 문제에 대한 기본적인 이해와 반복 실험을 거쳐서 사람이 직접 결정한다.

$$l = \frac{(a \times s_r + b)}{c^{d \times s_l + e}} \qquad (1)$$

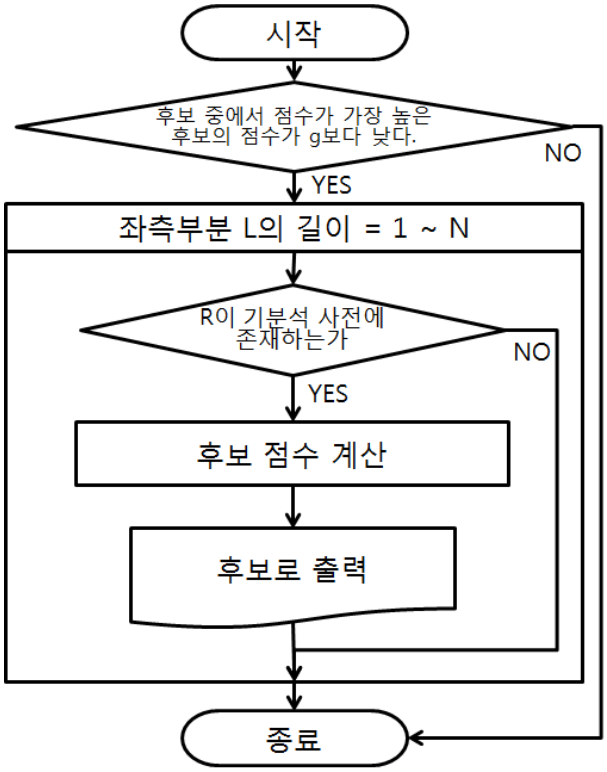
UTagger의 기분석 사전은 등록된 어절 또는 부분어절이 짧을수록 빈도가 매우 높아지는 성향을 보인다. 이 때문에 조사부분의 길이가 짧을수록 후보 점수가 높아질 수 있다. 따라서 다음과 같이 조사부분의 빈도를 수정하여 사용한다.

$$r' = r^f \qquad (2)$$

$r$은 우측부분인 조사부분의 빈도이며, $r'$은 수정된 빈도이다. 그리고 f는 a~e와 마찬가지로 상수이다.

### 3.3 미등록 명사 추정 알고리즘

<그림 1>은 상술한 3.1과 3.2를 종합하여 전체 알고리즘을 순서도로 표현한 것이다. 먼저 미등록 명사 추정을 시도할 것인지를 판단하기 위해서 기존의 후보 점수들 중에서 가장 높은 점수를 확인한다. 이 때 상수 g도 a~f와 같은 방법으로 결정한다. 미등록 추정을 하기로 결정하면 어절을 두 개로 나눈다. 이 때 나누는 방법은 어절의 길이만큼 존재하기 때문에 순환 처리를 한다. 좌측부분 L이 결정되면 어절의 나머지는 우측부분 R이 되고, 이것은 일반적으로 조사부분에 해당한다. 단, 여기서 구체적으로 R이 반드시 조사여야 한다고 엄격하게 제한하지 않는다. R은 등록된 명사가 될 수도 있으며 '이/VCP' 로 시작할 수도 있다. 여기서 중요한 것은 R이 어떤 형태로든 L과 조합이 가능하다면 후보로 생성될 수 있어야 한다는 것이다. 결국은 UTagger의 후보 점수 계산에서 L과 R의 조합 가능성을 계산할 것이기 때문에 R이 기분석 사전에 존재하는지만 먼저 판단한다. 이렇게 L과 R이 준비되면 후보 점수 계산을 수행하고 후보로 출력한다. 모든 순환이 끝나고 미등록 명사 추정 알고리즘이 종료되면 UTagger가 점수가 지나치게 낮은 후보들을 자동으로 삭제할 것이다.




<그림 1> 미등록 명사 추정 순서도

## 4. 실험 및 결과 활용

### 4.1 실험 결과

본 논문이 소개하는 미등록 명사 추정 알고리즘이 기존 형태소 분석기의 정확률을 향상시키는지 실험하기 위해서 기존 형태소 분석기의 정확률을 측정하는 방법을 그대로 사용하였다. UTagger는 세종말뭉치 990만 어절과 표준국어대사전을 학습하고 110만 어절을 실험하며, 이 때 미등록 명사 추정 알고리즘을 적용하여 추가로 후보를 생성하였다. 어절 단위로 정확률을 측정하였으며, 실험용 110만 어절에 나타난 명사 형태소 중에 대부분은 학습용 990만 어절에 나타나기 때문에 실제로는 미등록 명사는 많이 존재하지 않는다. 이런 실험 환경적 특성 때문에 본 논문이 소개하는 방법은 정확률에 큰 변화를 주지 못한다.

실험 결과는 <표 3>에 나타나있다. 기존 UTagger의 정확률에 비하여 약 0.01%의 정확률만 향상되었지만 미등

<표 3> 미등록 명사 추정 실험 결과

| 실험 종류 | 정확률 |
|---|---|
| 기존 UTagger | 96.3998% |
| 미등록 명사 추정 사용 | 96.4113% |

록 명사 추정을 사용함으로 인해서 특별히 처리 시간이 눈에 띄게 변하지 않았고 메모리 사용량에 변화도 없었기 때문에 충분히 의미가 있는 것으로 판단된다. 특히 이 실험은 미등록 명사가 등록되지 않은 어절에 대해서도 적용되었기 때문에, 만약 미등록 명사가 없는 어절에서 미등록 명사를 추정하게 되면 오히려 정확률이 낮아질 수 있게 된다는 점을 고려하면 이 실험 결과는 더욱 의미가 있다고 판단된다.

상술한 바와 같이 미등록 명사 추정을 통해서 기존에 정확히 태깅되던 어절이 오답으로 태깅될 수도 있다. 이런 경우를 측정한 결과가 <표 4>이다. 미등록 명사 추정을 적용하게 되어서 190개가 정답으로 변했으나 62개가 오답으로 변하여 대략 130개 어절만큼 긍정적인 효과를 주었다.

<표 4> 미등록 명사 추정을 적용하여 생긴 변화

| 변화 형태 | 어절 수 |
|---|---|
| 오답->정답 | 190개 |
| 정답->오답 | 62개 |

오답에서 정답으로 변한 어절들은 다음과 같다.

> 바잉오피스, 핸드캡, 컬럼니스트, 썬키스트, 플륨, 잔탁의, 리네이춰링, 여세부쟁, 비스므쓰, 스트립댄서, 얼짱, 퍼블리셔스 ...

대체로 외래어가 많은 것을 알 수 있다. 반대로 정답에서 오답으로 변한 어절들은 다음과 같으며 복합명사이거나 잘못 붙여 쓴 경우가 많았다.

> 준항모급, 버섯철, 빼어닮은, 청어잡이나, 미달금만큼, 놋천장처럼, 바꿔심기, 흰손, 클줄이야, 고른뒤, 기쁜소식, 재영토화된다, 흰별을 ...

수식에 사용된 상수들은 반복 실험 결과 식 (3)과 같이 결정되었으며, 임계치 g는 1로 결정되었다. 여기서 1은 UTagger의 점수 체계를 고려한다면 쉽게 이해될 수 있는 부분이다. UTagger는 세종말뭉치에서 어절 전체가 1번이라도 등장할 경우에 점수 대신에 빈도가 적용된다.

$$l = \frac{(1/16 \times s_r + 0.5)}{200^{2 \times s_l - 5}} \qquad (3)$$

$$r' = r^{0.075}$$

4.2 결과 활용 : 미등록 명사 반자동 수집기

3장에서 소개한 미등록 명사 추정 방법은 매우 소극적인 방법으로, 미등록 명사가 거의 발생하지 않는 실험 환경에서 기존 형태소 분석기에 부정적인 영향을 최소화하기 위해 설계된 것이다. 따라서 미등록 명사를 포함한 어절임에도 그대로 오답으로 태깅하는 경우도 있을 것으로 추측된다. 이 문제를 오직 자동화된 방법으로 해결하는 것에는 한계가 있을 것이기 때문에 사람이 직접 확인하면서 미등록 명사들을 수집하는 방법이 더욱 효과적일 것이다.

3장에서 소개한 방법에서 미등록 명사의 가능성을 확인하는 상수는 g이다. 이 임계치 g를 더 높이면 미등록 명사가 아닌 어절들도 더 많이 통과되겠지만 사람이 직접 태깅할 것을 고려한다면 훌륭한 반자동 수집기로 활용할 수 있을 것이다.

그러나 점수 임계치 g보다 점수 계산 방식을 일부 수정하는 것이 더 효과적일 것으로 판단된다. 미등록 명사들은 복합명사로 잘못 분석되는 경우가 많은데 이 경우에 점수 계산은 명사에서 명사로의 품사 전이를 고려하게 된다. 따라서 명사-명사 전이 확률을 일시적으로 낮추는 방식으로 전체 점수를 낮추면 임계치 g를 변경하는 것 보다 더 효과적으로 미등록 명사를 수집할 수 있을 것이다.

실제로 간단한 실험결과 g를 수정하기만 해서는 미등록 명사가 없는 어절들을 많이 수집하였고, 예를 들어 '아이패드'를 미등록 명사로 수집하지 못했다. 왜냐하면 '아이'와 '패드'는 각각 빈도가 높아서 후보의 점수가 높게 계산되었기 때문이다. 그러나 명사-명사 전이 확률을 낮추었을 때에는 쉽게 수집할 수 있었다.

## 5. 결론

본 논문은 미등록 명사가 거의 발생하지 않는 환경에서 미등록 명사 추정 알고리즘을 적용하여 긍정적인 결과가 나오는 방법을 소개하였다. 또한 사람의 확인 작업을 함께하여 미등록 명사들만 수집하는 방법을 제안하였다. 이런 방법들은 모두 기존 형태소 분석기에 대한 이해를 바탕으로 하고 있으며, 기존 형태소 분석기가 작동하기 위해 사용하는 컴퓨터 자원에 최소한의 변화만을 주도록 설계되었다.

실험 환경에서 미등록 명사가 거의 존재하지 않았기 때문에 정확률에 변화가 미미했으나, 그 변화 정도가 지나치게 미미했다고 추측되며, 실제로 미등록 명사가 얼마나 존재하는지를 정확히 알 수 없었기 때문에 본 논문이 소개하는 방법의 우수성을 검증하기 위해서는 새로운 실험 방법이 필요할 것이다. 또한 본 논문에서 소개하는 방법은 매우 간단한 방법이며, 기존의 미등록 명사 추정 연구에서 제안하는 기계학습 방법들과 적절히 병합한다면 컴퓨터 자원의 사용을 크게 늘리지 않으면서도 전체 성능을 향상시킬 수 있을 것으로 예상된다.

## 참고문헌

[1] 최맹식, 김학수, "기계학습에 기반한 한국어 미등록 형태소 인식 및 품사 태깅", 정보처리학회논문지, The KIPS transactions. Part B. 제18B권, 제1호, pp. 45-50, 2011.

[2] 김형철, 서형원, 김재훈, "접사 정보를 이용한 영어 미등록어의 품사부착 성능개선", 2009년도 한국마린

엔지니어링학회 공동학술대회 논문집, pp.375-376, 2009.
[3] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위 논문, 1993.
[4] 김선호, 윤준태, 송만석, "한국어 문서 처리를 위한 동적 생성로컬 사전 기반 미등록어 분석", 정보과학회논문지:소프트웨어 및 응용 제29권 제6호, pp.407-416, 2002.
[5] 박재한, 김명선, 노대욱, 나대열, "백오프 통계정보를 이용한 미등록어 포함 복합명사의 분해", 제16회 한글 및 한국어 정보처리 학술대회, 제16권, 제1호, pp. 65-72, 2004.
[6] 박봉래, 황영숙, 임해창, "용례 분석에 기반한 미등록어의 인식", 정보과학회논문지(B), 제25권, 제5호, pp. 397-407, 1998.
[7] 신준철, 옥철영, "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기", 한국정보과학회논문지 : 소프트웨어 및 응용, 제39권, 제5호, pp. 415-424, 2012.
[8] 신준철, 옥철영, "부분어절 조건부확률 기반 동형이의어 태깅 모델", 한국정보처리학회논문지, 소프트웨어 및 데이터 공학, 게재 예정, 2014.