DOI: 10.3745/KIPSTB.2011.18B.1.045

기계학습에 기반한 한국어 미등록 형태소 인식 및 품사 태깅

최 맹 식[†] · 김 학 수[††]

요 약

한국어 형태소 분석에서 미등록 형태소 오류들은 2가지 유형으로 나뉜다. 첫 번째 오류 유형은 형태소 분석기가 어떤 형태소열도 찾아내지 못하는 것이고, 두 번째 오류 유형은 등록 형태소들의 잘못된 조합을 찾아내는 것이다. 지금까지 대부분의 기존 미등록 형태소 추정 기술들은 단지 첫 번째 오류 유형에만 초점을 맞추어 왔다. 본 논문에서는 2가지 유형의 오류들의 모두 다룰 수 있는 미등록 형태소 추정 방법을 제안한다. 제안 방법은 SVM(Support Vector Machine)을 이용하여 미등록 형태소 오류들을 포함할 가능성이 있는 어절들을 검출한다. 그리고 CRFs(Conditional Random Fields)를 이용하여 검출된 어절들의 형태소 분리와 품사 태깅을 수행한다. 실험에서 제안 방법은 기능어 최장 일치 기반의 전형적인 방법보다 뛰어난 성능을 보였다. 실험 결과에 기초하여 미등록 형태소 오류의 두 번째 유형이 한국어 형태소 분석의 성능을 올리기 위해서 꼭 다루어져야 한다는 것을 알 수 있었다.

키워드 : 미등록 형태소 추정, 미등록 형태소 인식, 미등록 형태소 태깅

Part-Of-Speech Tagging and the Recognition of the Korean Unknown-words Based on Machine Learning

Maengsik Choi[†] · Harksoo Kim[††]

ABSTRACT

Unknown morpheme errors in Korean morphological analysis are divided into two types: The one is the errors that a morphological analyzer entirely fails to return any morpheme sequences, and the other is the errors that a morphological analyzer returns incorrect combinations of known morphemes. Most previous unknown morpheme estimation techniques have been focused on only the former errors. This paper proposes a unknown morpheme estimation method which can handle both of the unknown morpheme errors. The proposed method detects Eojeols (Korean spacing units) that may include unknown morpheme errors using SVM (Support Vector Machine). Then, using CRFs (Conditional Random Fields), it segments morphemes from the detected Eojeols and annotates the segmented morphemes with new POS tags. In the experiments, the proposed method outperformed the conventional method based on the longest matching of functional words. Based on the experimental results, we knew that the second type errors should be dealt with in order to increase the performance of Korean morphological analysis.

Keywords : Unknown Morpheme Estimation, Unknown Morpheme Recognition, Unknown Morpheme Tagging

1. 서 론

텍스트 문서로부터 정보를 추출하기 위해서는 형태소 분석과 품사 부착 단계가 중요하다. 형태소 분석은 사전을 기반으로 이루어지는데, 사전에 등재되어 있지 않은 어휘로 인해 형태소 분석 및 품사 부착 시에 성능이 크게 떨어진다. 특히 전문용어와 같은 신조어는 미리 사전에 포함되기가 어렵다. 이와 같이 미등록어는 형태소 분석 단계에서 형태소 사전에 등재되어 있지 않은 단어를 의미하며 이러한 단어로 인해 형태소 분석기의 성능이 크게 떨어진다.

한국어는 영어와 다르게 하나의 단어가 여러 형태소의 조합으로 이루어지게 된다. 그래서 미등록어의 대상 경계를 먼저 구분해야하며, 품사 추정에 사용할 수 있는 정보가 많지 않다. 따라서 대부분의 한국어 미등록어 추정 연구에서는 어절을 내용어와 기능어의 조합으로 보고, 미등록어의

※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2010-0009875).
† 준 회 원 : 강원대학교 컴퓨터정보통신공학과 석사과정
†† 정 회 원 : 강원대학교 컴퓨터정보통신공학전공 교수(교신저자)
논문접수 : 2010년 7월 28일
수 정 일 : 1차 2010년 10월 6일
심사완료 : 2010년 10월 20일

내용어가 대부분 명사라는 점을 이용하여 어절을 두 부분으로 나누어 미등록어 대상을 찾고, 명사로 추정하는 방법이 일반적이다. 그러나 이런 방법은 어절을 구성하는 형태소가 3개 이상일 경우에 내용어와 기능어 두 개로 구분하므로 명확한 구분을 하기가 어려우며, 미등록어의 대상이 되는 어휘가 명사가 아닐 경우에는 처리가 어렵다. 또한 미등록어 처리 대상은 형태소 분석에 실패한 어절을 대상으로 이루어지므로 미등록어를 포함하는 어절이 잘못된 형태소의 조합으로 분석된다면 어떠한 처리도 할 수 없다는 문제가 있다.

본 논문에서는 미등록어 처리 문제를 형태소 분석기의 후처리를 통해 처리함으로써, 미등록어를 포함하는 어절의 품사부착 결과가 잘못된 것을 인식하여 품사 재부착을 시도한다. 본 논문의 구성은 다음과 같다. 먼저, 2장에서 관련연구에 대해 알아보고 3장에서 미등록어 추정을 위한 방법을 제안한다. 4장에서 실험 및 평가, 5장에서 결론을 맺는다.

2. 관련 연구

영어 미등록어 추정의 경우, 하나의 단어가 하나의 품사로 구성되므로 형태소 분석 단계가 필요 없다. 그러므로 미등록어 추정 시 단어 내에서 미등록어의 경계를 인식할 필요 없이 해당 단어의 품사를 추정한다. 영어 단어의 품사 추정에는 해당 단어 주변의 문맥 자질, 또는 해당 단어의 접두사, 접미사 그리고 대문자로 시작하는 정보 등을 이용할 수 있다. 국내의 영어 미등록어 추정에 관한 연구로 김형철 외[1]가 있다. 김형철 외[1]는 접두사, 접미사 그리고 이전 품사 자질을 CRFs(Conditional Random Fields)를 이용하여 미등록어 품사를 추정하였다.

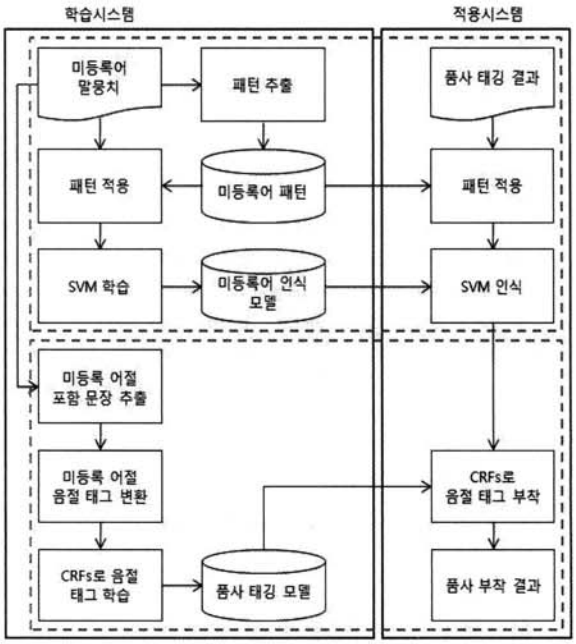
한국어 미등록어 추정의 경우에는 품사 추정 이전에 어절을 이루는 형태소들의 경계를 구분하여야 한다. 미등록어의 경우, 사전에 없는 어휘이므로 경계 구분이 어렵다. 따라서 대부분의 미등록어 추정 연구에서는 어절을 내용어와 기능어의 조합으로 보고 문제를 해결하고 있다. 가장 일반적인 방법은 미등록 어절을 명사와 조사의 결합으로 보고, 어절의 오른쪽에서 최장 조사를 떼어내고 남은 부분을 미등록 명사로 인식하는 최장조사 분리 방법이다[2]. 박봉래 외[3]는 동일한 미등록어가 포함된 둘 이상의 유사 어절들을 분석하여 내용어와 기능어를 추출하고 동일한 내용어에 붙는 기능어들의 종류에 따라 미등록어를 인식하였다. 김선호 외[4]는 각 문서를 대상으로 형태소 분석 단계 이전에 미등록어의 대상이 되는 단어들을 동적 사전으로 만들어 형태소 분석 시에 사용하였다. 기존 연구들의 경우, 미등록어의 대상을 체언으로 한정하는 점, 어절을 두 형태소의 구성으로 보고 미등록어는 어절의 앞부분에 하나만 존재할 경우를 대상으로 하였다. 또한 미등록어로 인해 형태소 분석이 되지 않는 어절을 대상으로 하였다.

본 연구에서는 형태소 분석 실패를 통하여 미등록어라고 판명된 대상의 품사 부착문제 이외에도 미등록어로 인한 형태소 분석 오류 어절을 인식하여 품사 재부착을 하는 방법에 대해 논의한다.

3. 미등록어 인식 및 품사 태깅

3.1 시스템 개요

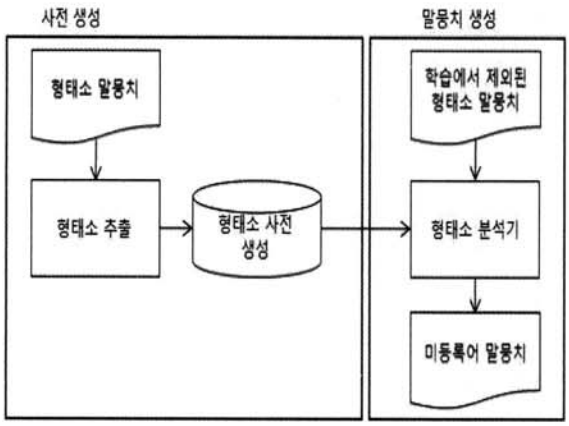
제안 시스템은 (그림 1)과 같이 형태소 분석 결과에서 미등록어로 인해 형태소 분석이 잘못된 어절(이하 미등록 어절)을 인식하는 부분과 미등록 어절에 다시 품사를 부착(이하 미등록 어절 품사 태깅)하는 부분으로 나눌 수 있다. 미등록 어절 인식의 경우에 올바른 어절을 잘못 인식하는 것을 막기 위하여 재현율은 다소 떨어지더라도 정확률이 높은 방법을 사용해야 한다. 이를 위해서 본 논문에서는 미등록 어절 패턴과 SVM(Support Vector Machine)[5]을 단계적으로 사용하는 방법을 제안한다. 그리고 미등록 어절 품사 태깅은 레이블링(labeling) 문제에서 좋은 성능을 보이고 있는 CRFs[6]를 이용하는 방법을 제안한다.



(그림 1) 제안 시스템 구조도

제안 시스템을 학습하기 위한 말뭉치는 다음과 같이 3가지로 구성된다. 첫째는 '대학/명사+생선/명사+교회/명사+가/조사'와 같이 '대학생선교회'라는 미등록어로 인해 '대학', '생선', '교회'라는 잘못된 등록어들의 조합으로 품사가 태깅된 어절(이하 NA⁻ 어절;Not-Analyzed Negative)들이 포함된 문장이다. 둘째는 '안드로이드/분석불능범주+는/조사'와 같이 '안드로이드'라는 미등록어로 인하여 형태소 분석에 실패한

어절(이하 NA^+ 어절;Not-Analyzed Positive)들을 포함하는 문장이다. 그리고 셋째는 올바르게 형태소 분석된 어절들만이 포함된 문장이다. 이러한 학습 말뭉치를 자동으로 만들기 위해서 본 논문에서는 (그림 2)와 같은 과정을 거친다. 먼저, 기존의 형태소 말뭉치[1]를 9:1의 비율로 사전 생성용 데이터와 말뭉치 생성용 데이터로 나눈다. 그리고 사전 생성용 데이터에서 형태소들을 추출하여 형태소 사전을 만든다. 다음으로 새롭게 만들어진 형태소 사전을 형태소 분석기에 탑재하여 말뭉치 생성용 데이터를 형태소 분석한다. 이렇게 하면 형태소 사전에 포함되지 않은 형태소를 포함하는 어절들은 분석이 되지 않거나(NA^+ 어절로 분석되거나), 다른 형태소들의 조합으로 분석(NA^- 어절로 분석) 되기 때문에 상기한 3가지 종류의 학습 말뭉치를 얻을 수 있다.



(그림 2) 학습 말뭉치 생성 구조도

3.2 미등록어 인식

기존의 미등록어 추정 시스템들은 주로 NA^+ 어절만을 대상으로 하지만 제안 시스템은 NA^- 어절도 대상으로 한다. 학습 말뭉치에서 형태소 분석을 실패한 어절 수와 다른 형태소의 조합으로 분석된 어절 수를 살펴보면 NA^+는 8,185 어절, NA^-는 23,791 어절로 NA^- 어절이 전체의 74.4%에 해당한다. 이것은 NA^- 어절을 인식하고 형태소 분석 오류를 수정하는 것이 매우 중요하다는 것을 보여준다고 할 수 있다. 본 논문의 미등록어 인식 대상은 NA^- 어절이다. NA^+ 어절은 형태소 분석기의 일반적인 출력 결과에서 알 수 있기 때문에 본 논문에서는 미등록어 인식 대상으로 삼지 않는다.

NA^- 어절을 인식할 때, 올바르게 형태소 분석된 어절을 오인식하지 않는 것이 매우 중요하기 때문에 본 논문에서는 2단계 방법을 이용한다. 1단계로 학습 말뭉치에서 NA^- 어절 패턴을 추출한다. (그림 3)은 NA^- 어절 패턴의 예이다. (그림 3)에서 영문은 21세기 세종계획[8]에서 정의한 품사 태그(tag)를 의미하고, 품사 태그 앞에 붙은 숫자는 음절의 길이를 의미한다. 즉, '1NNP-1NNP-1VCP-1EC'는 품사가 'NNP'인 한 음절 형태소, 품사가 'NNP'인 한 음절 형태소, 품사가 'VCP'인 한 음절 형태소, 그리고 품사가 'EC'인 한 음절 형태소로 이루어진 NA^- 어절 패턴을 의미한다.

```
1NNP-1NNP-1VCP-1EC
1XPN-1NNG-1XSV-1EC
1XPN-1NNG-1NNG-1NNG-1JKO
```

(그림 3) NA^- 어절 패턴의 예

학습 말뭉치를 대상으로 NA^- 어절 패턴을 생성한 후, 수식 (1)을 이용하여 패턴별 신뢰도를 계산한다[7].

$$Score(Rule)=\frac{\frac{NA^-(Rule)}{NA^-(Rule)+CA(Rule)}\times\log_2(NA^-(Rule)+1)}{MaxScore} \quad (1)$$

수식 (1)에서 NA^-는 패턴이 NA^- 어절에 나타난 수이고, *CA(Correctly Analyzed)*는 패턴이 올바르게 형태소 분석된 어절에서 나타난 수이다. *MaxScore*는 각 패턴들에 부여된 신뢰도 중에서 가장 큰 값으로 패턴 신뢰도를 정규화하는데 사용된다. 계산된 신뢰도를 바탕으로 내림차순 정렬하여 일정 점수 이상의 패턴을 선택하는 방법으로 NA^- 어절 패턴을 생성한다. 본 논문에서는 0.14 이상의 신뢰도를 가지는 패턴을 사용하였다. NA^- 어절 패턴 생성이 끝나면, NA^- 어절 패턴과 매칭(matching)된 어절들만을 SVM의 입력으로 하여 NA^- 어절인지 여부를 최종 결정한다. SVM 학습을 위한 자질로는 <표 1>과 같이 어절 내 형태소 바이그램과 가장 긴 형태소의 음절 길이, 그리고 어절 내 체언 형태소의 수를 사용하였다.

〈표 1〉 SVM 입력 자질

자질	설명	예
형태소 바이그램	어절의 형태소 분석 결과에서 형태소 바이그램	<S>_이, 이_운총, 운총_이, 이_라고, 라고_<S>
음절 길이	어절의 형태소 분석 결과에서 가장 긴 음절의 길이	2
체언 수	어절의 형태소 분석 결과에서 체언으로 분석된 형태소의 수	2

3.3 미등록어 품사 태깅

미등록어 인식이 끝나면 NA^- 어절과 NA^+ 어절 모두를 대상으로 품사 태깅을 다시 수행한다. 품사 재태깅은 <표 2>와 같이 형태소 경계인식을 위한 정보와 품사 정보가 결합된 태그를 사용한다.

1) 본 논문에서는 21세기 세종계획[8] 형태소 말뭉치 139,757문장을 사용

〈표 2〉 품사 재부착용 음절 단위 태그 셋

태그	설명
BX	품사가 X인 형태소 시작 음절
IX	품사가 X인 형태소 중간 음절
EX	품사가 X인 형태소 끝 음절
SX	품사가 X인 한 음절 형태소

예를 들어, '희/NNG+끗/NNG+하/XSA+ㄴ/ETM'이라는 어절이 NA^- 어절로 인식되었다면, 품사 재태깅은 '희끗하ㄴ'이라는 문자열에 '희/BXR+끗/EXR+하/SXSA+ㄴ/SETM'과 같이 음절('ㄴ'과 같이 자소로 분리된 문자는 그 자체를 음절로 간주함) 단위 태그 셋을 부착하는 것이다. 음절 단위 품사 태깅을 위하여 본 논문에서는 CRFs를 사용한다. (그림 4)는 '희끗한'이라는 어절에 대한 CRFs 학습 패턴의 예이다.

```
RT:VXEC FT:XSNJKS P-1:<S> P0:희 P+1:끗 T:1 BXR
RT:VXEC FT:XSNJKS P-1:희 P0:끗 P+1:하 T:1 EXR
RT:VXEC FT:XSNJKS P-1:끗 P0:하 P+1:ㄴ T:1 SXSA
RT:VXEC FT:XSNJKS P-1:하 P0:ㄴ P+1:<S> T:1 SETM
```

(그림 4) CRFs 학습 패턴의 예

(그림 4)에서 'RT'는 미등록어 앞 어절의 대표 품사이다. 대표 품사는 어절을 이루는 형태소 중 내용어의 마지막 형태소 품사와 기능어의 마지막 형태소 품사 쌍이다. 'FT'는 미등록어 뒷 어절의 대표 품사이다. 'P-1'은 현재 음절의 바로 앞 음절, 'P0'은 현재 음절, 'P+1'은 현재 음절의 바로 뒷 음절이다. 'T'는 현재 음절의 유형(1:한글, 2:한자, 3:알파벳, 4:숫자, 5:기타)이다. 마지막은 정답 품사이다. 미등록 어절에 대해서 CRFs를 이용하여 음절 태그를 부착한 후, <표 3>과 같이 일반적인 형태소 단위 품사 태그로 변환한다. 음절 태그를 형태소 단위 품사 태그로 변환하기 위해서는 먼저 'SX' 음절을 형태소로 생성하거나, 'BX'부터 'EX'까지 음절을 결합하여 형태소를 생성한다. 그리고 'BX'부터 'EX'까지의 품사 정보 'X' 중에서 가장 많은 수를 차지하는 것을 해당 형태소의 품사로 결정한다.

〈표 3〉 형태소 단위 품사 태그 부착의 예

음절 태그	형태소 단위 품사 태그
희/BXR+끗/EXR+하/SXSA+ㄴ/SETM	희끗/XR+하/XSA+ㄴ/ETM
마/BNNG+이/INNG+크/ENNG+로/SJKO	마이크/NNG+로/JKO

4. 실험 및 평가

4.1 미등록어 인식 모델 실험

학습 및 실험에는 21세기 세종계획[8] 형태소 말뭉치 중 랜덤(random)하게 샘플링(sampling)한 139,757문장을 10배 교차 검증(10-fold cross validation) 방법으로 실험하였다. 미등록어 인식 모델의 성능 평가를 위해 수식 (2)의 정밀도(accuracy), 수식 (3)의 재현율(recall), 수식 (4)의 정확률(precision)을 측정하였다.

$$\text{정밀도} = \frac{\text{올바르게 인식한 어절 수}}{\text{전체 어절 수}} \quad (2)$$

$$\text{재현율} = \frac{\text{인식한 정답 } NA^- \text{ 어절 수}}{\text{전체 } NA^- \text{ 어절 수}} \quad (3)$$

$$\text{정확률} = \frac{\text{인식한 정답 } NA^- \text{ 어절 수}}{NA^- \text{로 인식한 어절 수}} \quad (4)$$

<표 4>는 어절단위 미등록어 인식 모델의 성능이다. NA^+ 어절은 미등록어 인식 대상이 아니므로 성능에서 제외하였다.

〈표 4〉 미등록어 인식 성능

	정밀도	재현율	정확률
평균	96.84	8.77	83.70

미등록어 인식 결과 중 재현율이 상당히 낮게 나온 이유는 오인식률을 낮추기 위해서 정확률을 높였기 때문이다. 미등록어 추정의 목적이 잘못된 형태소 분석 결과를 올바르게 고치는 것이기 때문에 재현율이 다소 낮더라도 오인식되는 경우를 막는 것이 가장 중요하며, 이를 위하여 본 논문에서는 정확률을 높이는 방향으로 실험을 진행하였다. <표 5>는 미등록어 오인식률을 보여준다.

〈표 5〉 미등록어 인식 결과

정답 \ 시스템	NA^- 어절	올바른 어절
NA^- 어절	2,088	21,704
올바른 어절	407	1,982,217

4.2 미등록어 태깅 모델 실험

미등록어 태깅 모델의 성능 평가를 위해 수식 (5)와 같이 정확률을 측정하였다.

$$\text{정확률} = \frac{\text{정답품사로 태깅된 형태소 수}}{\text{태깅된 형태소 수}} \quad (5)$$

미등록어 태깅 모델의 성능은 NA^+ 어절과 NA^- 어절만을 대상으로 형태소 단위로 측정하였다. 미등록어 태깅 모델의

정확률은 평균 75.18%였으며, CRFs를 이용하는 기존 영어 미등록어 추정 모델[1]의 성능 83.77%보다는 낮았다. 제안 시스템은 형태소의 경계 구분까지 포함된 성능으로써 직접적인 비교는 어렵다고 생각된다. 또한 한국어 기존 시스템들은 체언(주로 고유명사)만을 대상으로 실험을 한 것이며, 제안 시스템은 세종 품사 태그 셋 42개(NA, NF, NV제외)를 대상으로 실험한 것이기 때문에 비교 자체가 무의미하다고 생각된다. 미등록어 오류의 유형에는 체언 이외에도 동사와 부사 등의 미등록어가 존재하였고, 체언 중에서도 고유명사 보다 일반명사의 비율이 높았다. 이러한 사실은 본 논문에서 제안하는 것과 같이 모든 품사를 대상으로 하는 미등록어 추정에 관한 연구가 필요하다는 것을 말해준다.

4.3 통합 모델 실험 (미등록어 인식 + 미등록어 태깅)

<표 6>은 미등록어 인식 모델에 의해 인식된 NA^- 어절과 형태소 분석에 실패한 NA^+ 어절을 대상으로 품사 태깅한 결과를 대표적인 미등록어 추정 방법인 조사 최장일치 모델과 비교한 것이다. <표 6>에서 보는 것과 같이 제안 모델은 NA^+ 어절만을 대상으로 비교해도 조사 최장일치 모델보다 월등히 높은 성능을 보였다. 또한 대부분의 기존 미등록어 추정 모델들[1, 3]은 조사 최장일치 모델과 같이 NA^- 어절을 대상으로 하고 있지 않다. 그러나 제안 모델은 인식 재현율이 낮음에도 불구하고 일부 NA^- 어절을 처리할 수 있음을 보여주었다.

〈표 6〉 미등록어 추정 결과

모델	인식 대상	대상 어절수	인식 어절수	올바르게 품사 태깅된 어절수
조사 최장일치 모델	NA^-	23,791	-	-
	NA^+	8,135	-	814
제안 통합 모델	NA^-	23,791	2,088	1,355
	NA^+	8,135	-	3,634

<표 7>은 21세기 세종계획 형태소 말뭉치 중 139,757문장을 대상으로 미등록어 처리를 하지 않았을 때와 최장조사 일치로 미등록어를 처리하였을 때, 기존 형태소 분석기의 성능이 어느 정도 향상될 수 있는지를 보여준다.

〈표 7〉 형태소 분석 후처리 결과

	미등록어 처리 안함	조사 최장일치 모델	제안 통합 모델
평균	94.86	94.87	95.21

5. 결론 및 향후 과제

본 논문에서는 한국어 미등록어 추정을 위하여 기계학습을 기반으로 형태소 경계를 인식하고 품사를 재부착하는 모델을 제안하였다. 제안 모델은 기존의 모델들과는 달리 미등록어로 인해 형태소 분석이 실패한 어절뿐만 아니라 잘못 형태소 분석된 어절도 처리할 수 있다는 장점이 있다. 또한 조사 접속과 같은 간단한 휴리스틱으로 미등록어 어절을 내용어와 기능어로 나누는 것이 아니라 기계학습을 기반으로 형태소 경계를 인식하고 품사를 추정함으로써 보다 안정된 성능을 기대할 수 있다는 장점이 있다. 향후에는 미등록어 어절 인식을 위한 좀 더 좋은 자질을 발굴하여 정확률을 많이 떨어뜨리지 않으면서 재현율을 높이는 방법을 연구할 예정이다.

참 고 문 헌

[1] 김형철, 서형원, 김재훈, "접사 정보를 이용한 영어 미등록어의 품사부착 성능개선", 2009년도 한국마린엔지니어링학회 공동학술대회 논문집, pp.375-376, 2009.

[2] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위 논문, 1993.

[3] 박봉래, 황영숙, 임해창, "유사 어절의 TAIL 패턴 분석에 기반한 미등록 명사 추정", 1996년도 한국정보과학회 봄 학술발표 논문집 제23권 제1호, pp.907-910, 1996.

[4] 김선호, 윤준태, 송만석, "한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석", 정보과학회논문지:소프트웨어 및 응용 제29권 제6호, pp.407-416, 2002.

[5] Chang, C.-C. and C.-J. Lin., "LIBSVM: a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2001.

[6] McCallum, Andrew Kachites., "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu. 2002.

[7] Riloff, E., Jones, R., "Learning dictionaries for information extraction by multi-level bootstrapping," In Proceedings of the 16th National Conference on Artificial Intelligence, pp.474-479, 1999.

[8] http://www.sejong.or.kr (2010. 7. 5 방문).

최 맹 식

e-mail : nlpmschoi@kangwon.ac.kr

2009년 강원대학교 컴퓨터정보통신공학전공(공학사)

2009년~현 재 강원대학교 컴퓨터정보통신공학과 석사과정

관심분야 : 형태소 분석, 구문 분석, 관계 추출

김 학 수

e-mail : nlpdrkim@kangwon.ac.kr

1996년 건국대학교 전자계산학과(공학사)

1998년 서강대학교 컴퓨터학과(공학석사)

2003년 서강대학교 컴퓨터학과(공학박사)

2004년~2005년 CIIR in UMass, Amherst(박사후연구원)

2005년~2006년 한국전자통신연구원(선임연구원)

2006년~현 재 강원대학교 컴퓨터정보통신공학전공 교수

관심분야 : 자연어처리, 대화시스템. 정보검색, 질의응답시스템