http://dx.doi.org/10.3745/KIPSTB.2012.19B.3.195

NB 모델을 이용한 형태소 복원

김 재 훈† · 전 길 호††

요 약

한국어는 교착어이어서 형태소 분석 없이 품사 부착이 어려울 뿐 아니라 형태소를 분석할 때 다양한 어형 변화가 복원되어야 한다. 이것은 한국어 형태소 분석의 고질적인 문제 중 하나이며, 주로 규칙을 이용해서 해결한다. 규칙을 이용할 경우 주어진 문맥에 가장 적합한 복원을 어려워 여러 형태의 모호성을 생성하며, 이는 품사 부착에 의해서 해결된다. 본 논문에서는 이 문제를 기계학습 방법(Naïve Bayes 모델)을 이용하여 해결한다. 기계학습 모델의 입력 자질은 어형 변화가 발생하는 주변 음절이며 출력 범주는 복원된 음절이다. ETRI 구문 말뭉치를 이용한 실험에서 제안된 형태소 복원 모델을 사용한 형태소 단위의 품사 부착 성능은 97.5%의 F_1 점수를 보였으며 이 모델이 형태소 복원에 매우 유용함을 알 수 있었다.

키워드 : 품사 부착, NB 모델, 형태소 복원, 형태소 분석

Morpheme Recovery Based on Naïve Bayes Model

Jae-Hoon Kim† · Kil-Ho Jeon††

ABSTRACT

In Korean, spelling change in various forms must be recovered into base forms in morphological analysis as well as part-of-speech (POS) tagging is difficult without morphological analysis because Korean is agglutinative. This is one of notorious problems in Korean morphological analysis and has been solved by morpheme recovery rules, which generate morphological ambiguity resolved by POS tagging. In this paper, we propose a morpheme recovery scheme based on machine learning methods like Naïve Bayes models. Input features of the models are the surrounding context of the syllable which the spelling change is occurred and categories of the models are the recovered syllables. The POS tagging system with the proposed model has demonstrated the F_1-score of 97.5% for the ETRI tree-tagged corpus. Thus it can be decided that the proposed model is very useful to handle morpheme recovery in Korean.

Keywords : POS Tagging, Naive Bayes Model, Morpheme Recovery, Morphological Recovery

1. 서 론

품사 부착은 주어진 문장에서 각 단어가 가질 수 있는 가장 적합한 품사를 결정하는 것이며, 품사 부착 방법은 1990년대 이후 매우 다양한 방법으로 연구되어 왔다[1-2]. 말뭉치가 늘어나고 기계학습 기술이 발전하면서 매우 다양한 방법으로 품사 부착 문제를 해결했다. 그러나 한국어의 경우에는 다양한 기계학습 방법으로 그대로 적용할 수 없어서 주로 확률기반 방법[3]과 혼합 방법[4]에 치중되어 왔다. 왜냐하면 한국어는 교착어이므로 형태소와 형태소가 결합할 때 형태소 원형이 변형되어 형태소 분석이 없이는 형태소의 경계와 형태소의 원형을 정확히 분석할 수 없기 때문이다. 이와 같은 어려움에도 [5]에서는 사례 기반 학습 방법을 이용해서 한국어 품사 부착을 시도하였다. 그러나 [5]에서는 형태소의 원형이 복원되지 않고 단순히 형태소의 경계만 분리하고 분리된 형태소에 품사를 부착하는 방법이다. [5]에서 발생되는 문제를 해결하기 위해서 [6]에서는 규칙을 이용해서 형태소 복원을 시도하였다. 그러나 [6]에서 제안한 모델은 몇 가지 문제를 가지고 있다. 첫 번째 문제는 복합명사를 분석할 수 없다는 것이다. 예를 들어 “경제발전이”라는 어절을 분석하면 “경제발전/NN+이/JO[1)]”와 같이 분석되며 “경제/NN+발전/NN+이/JO”로 분석할 수 없다. 왜냐하면 음절 단위로 품사를 부착하고 부착된 음절품사

※ 이 논문은 한국전자통신연구원 위탁과제 “자동번역 지식 구축 도구에 대한 연구”로부터 일부 지원되었음.
† 종신회원 : 한국해양대학교 IT공학부 교수
†† 준 회 원 : 한국해양대학교 컴퓨터공학과 공학석사
논문접수 : 2012년 2월 21일
수 정 일 : 1차 2012년 3월 22일
심사완료 : 2012년 4월 11일

1) [6]에서 사용한 품사 NN은 명사이고 JO는 조사, VV는 동사, EM은 어미, SV는 접미사, EP는 선어말어미이다.

(syllable tag)가 같은 음절은 하나로 묶어서 하나의 형태소를 구성하기 때문이다. 두 번째 문제는 형태소 복원 모델이 완전하지 않다는 것이다. 왜냐하면 형태소 복원 규칙이 모호하기 때문이다. 예를 들면 “소리를 들어 보다”에서 “들어”를 분석할 때 “들/VV : 듣”이라는 형태소 복원 규칙을 사용하여 “듣/VV+어/EM”으로 분석된다. 이 경우는 올바른 분석이 되지만 “책을 들어 보다”에서 “들어”를 분석할 때도 완전히 같은 문맥이므로 같은 규칙이 적용되어 “듣/VV+어/EM”으로 분석된다. 이 경우의 올바른 분석은 “들/VV+어/EM”이며 그릇된 결과를 출력한다. 본 논문에서 첫 번째 문제를 해결하기 위해서 음절 단위로 품사를 부착할 때 [5]에서와 같은 방법으로 형태소가 분리되는 마지막 음절에는 “+”를 추가하였다. 두 번째 문제를 해결하기 위해서 본 논문에서는 기계학습 방법(Naïve Bayes(NB) 모델)을 이용해서 형태소 복원 모델을 제안했다. 그 결과 기존의 품사 부착 모델과 별다른 성능 차이를 보이지 않았으며 형태소 분석을 사용하지 않고도 품사 부착이 가능한 한국어 품사 부착 모델을 제안할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구로 기계학습을 이용한 한국어 품사 부착 방법에 대해서 간략히 살펴본다. 3장에서 음절 단위의 품사 부착 말뭉치의 부호화(coding scheme)에 대해서 기술하고 4장에서는 NB 모델를 이용한 형태소 복원 모델을 기술한다. 5장에서 제안된 모델은 실험하고 분석하고 6장에서 결론을 맺고자 한다.

2. 관련 연구: 기계학습을 이용한 한국어 품사 부착

본 절에서는 일반적인 기계학습 모델에 관해서 간단히 기술하고, 확률 기반 방법을 제외한 기계학습 방법을 이용한 기계학습 기반 한국어 품사 부착 방법에 대해서 기술한다.

일반적으로 기계학습 방법을 분류하면 매우 다양하며[7] 본 절에서는 지도학습 기반의 분류 모델(classification model)에 대해서 살펴본다. 분류 모델은 식 $y=f(\vec{x})$로 표현되고, 여기서 y는 범주(category, class)이고, $\vec{x}$는 자질 벡터(feature vector)이며, $f(\cdot)$는 분류기(classifier)이다. 다시 말해서 자질 벡터를 입력으로 받아서 특정 범주를 출력하는 시스템이며 분류기 $f(\cdot)$는 일반적으로 학습 데이터에 의해서 자동으로 생성되며 학습 데이터는 전문가에 의해서 구축된다(지도기반 학습). 이와 같은 모델은 품사 부착[8] 뿐 아니라 구문분석[9] 등 자연언어처리 전반에 사용되고 있다. 한국어 품사 부착의 경우에는 확률 모델을 이용한 기계학습 방법들[3-4]이 주로 사용되었으나 그 외의 방법들은 널리 사용되지 않았다.

[5]에서는 형태소 분석기를 사용하지 않고 사례기반 학습(instance-based learning)을 통해 단어를 분리하고, 분리된 단어의 범주를 결정하는 방법을 제안했다. 이 시스템의 입력은 음절의 자질 벡터이고 출력은 음절품사(syllable tag)[2]와 형태소 경계를 표시하는 ‘+’ 기호의 조합이다. 자질 벡터는 주어진 음절의 문맥 정보로서 주변 음절 정보, 받침 정보, 품사 정보로 구성되어 있다. 음절 기반의 품사 부착 말뭉치의 구축 방법에 대한 자세한 내용은 [5]를 참조하기 바란다. 이 논문은 처음으로 기계학습 방법을 이용한 한국어 품사 부착을 시도했다는 데는 큰 의의가 있으나, 형태소가 복원되지 않아서 일반적인 자연언어 처리 시스템에서 그대로 사용하는 데는 다소 무리가 있었다.

[6]은 기본적인 시스템을 구조는 [5]와 같다. 가장 큰 차이는 [5]에서 야기된 형태소 복원 문제를 규칙을 이용해서 해결한다는 것이다. 또 다른 차이로는 음절품사 부호화에 있다. 이 논문은 주어진 문장에서 모든 공백문자를 제거하고 대신에 어절을 BIO 부호화 방법[10]으로 구별한다. 예를 들어 어절 “할 수 있다.”를 BIO 부호화하면 “할/B 수/B 있/B 다/I ./I”가 된다. 또한 이 논문은 음절을 부호화할 때 한 음절이 두 형태소의 경계에 있다면 그 두 형태소가 가지는 모든 품사로 음절품사를 표현하였다. 예를 들면 어절 ‘제안했다’에 대해서 살펴보자. 이 어절의 품사 부착 결과는 “제안/NN 하/SV 었/EP 다/EM”이며 음절 ‘했’이 두 형태소 ‘하’와 ‘었’에 포함된다. 따라서 이 어절에 대한 음절품사 부호화 결과는 “제/NN 안/NN + 했/SVEP 다/EM”이다. 여기서 ‘SVEP’은 품사 ‘SV’와 ‘EP’가 결합된 복합품사이다. 이 논문은 CRF++을 이용해서 구현되었으며 96.31%의 정확도를 보여 확률 기반 모델과 비슷한 성능을 보여 충분히 실용적으로 사용할 수 있음을 보였다. 그러나 1장에서 언급했듯이 복합명사를 분석할 수 없는 복합명사 미분석 문제와 형태소 복원 모델이 완전하지 않은 형태소 복원의 불완전성 문제를 가지고 있다.

본 논문에서는 이들 두 문제를 해결하기 위해 음절 품사의 부호화 방법과 기계학습을 이용한 형태소 복원 모델을 제안한다.

3. 음절품사 부착 말뭉치의 부호화 및 생성

본 논문에서는 기본적인 음절품사 부호화 방법은 [5]와 같으나 [6]에서 제안한 복합품사 개념을 부가하였다(<표 1>에서 “V+E”[3]). <표 1>에서 “추억상자”가 복합명사이며 음절품사 부착 결과는 “추/N 억/N+ 상/N 자/N+”와 같다. 따라서 [6]에서 야기된 복합명사 미분석 문제는 본 논문에서 제안한 부호화 방법으로 자연스럽게 해결될 수 있다.

음절품사 부착 말뭉치는 형태소 품사 부착 말뭉치로부터 생성된다. <표 1>에서 볼 수 있듯이 한 표층 음절(surface syllable)(‘다’)이 여러 음절(‘이+다’)로 대응되는 경우도 있고, 경우에 따라서는 이와 반대의 경우도 발생되기 때문에 형태소와 표층 음절을 정렬시키는 것은 단순한 문제가 아니다.

2) 음절품사는 표층 음절이 포함된 형태소의 품사를 말한다. 예를 들면 어절 “학교가”가 가지는 표층 음절 ‘학’, ‘교’, ‘가’가 있고 각 음절의 품사는 각각 명사, 명사, 조사이다.

3) 본 논문에서는 [5]에서와 같이 단순화된 태그를 사용하며 ‘V’는 용언(동사, 형용사), ‘E’는 어미, ‘N’은 명사, ‘C’는 지정사, ‘SP’는 공백문자, ‘.’은 문장부호를 나타낸다.

본 논문에서는 사용하는 기본적인 정렬 알고리즘은 “최소교정거리를 이용한 어절(표층 음절열)과 형태소들의 정렬 알고리즘”[5]이며, 복합품사를 위해서 개선하여 사용하였다.

〈표 1〉 음절품사 부착 말뭉치의 부호화

어절	아름다운				ㅗ	추억상자다.						
형태소	아름답			ㄴ	ㅗ	추억		상자		이	다	.
품사	V			E	SP	N		N		C	E	.
음절	아	름	다	운	ㅗ	추	억	상	자	다		.
음절 품사	V	V	V	V+E	SP	N	N+	N	N+	C+E		.

4. NB 모델을 이용한 형태소 복원

본 절은 음절품사 부착을 기반으로 하는 한국어 형태소 복원에 관하여 기술한다. 대부분의 한국어 어절은 두 개 이상의 형태소로 구성되어 있으며 대부분은 단순한 두 형태소의 결합으로 이루어진다. 그러나 동사나 형용사가 어미와 결합할 경우에는 다양한 형태 변이가 발생되므로 형태소 분석에서는 이를 원래의 형태소로 복원해야 한다. 이를 형태소 복원(morpheme recovery)이라고 한다. 예를 들어 형용사 ‘아름답다’와 어미 ‘ㄴ’이 결합하면 어절 ‘아름다운’이 되므로 표층형으로 나타나는 ‘아름다운’은 “아름답+ㄴ”로 분석되어야 한다. 본 논문에서 형태소 복원 모델은 Naïve Bayes(NB) 모델이며 (그림 1)과 같이 그림(graphical model)으로 표현할 수 있다[4]. 식 (1)은 (그림 1)과 같은 모델의 NB 분류기이다.

$$\hat{c} = \arg\max_{c \in C} P(c)\prod_{j=1}^{9} P(x_j|c)$$
$$= \arg\max_{c \in C} \log P(c) + \sum_{j=1}^{9} \log P(x_j|c) \qquad (1)$$

여기서 c는 범주를 나타내고, C는 범주 집합({‘답’, ‘ㄴ’, ‘이+다’, ... })이다. 또한 $\hat{c}$는 여러 개의 범주 중에서 가장 높은 확률값을 가지는 범주 c이며, x_j는 j번째 자질을 나타내며, 자질집합 X는 식 (2)와 같다.

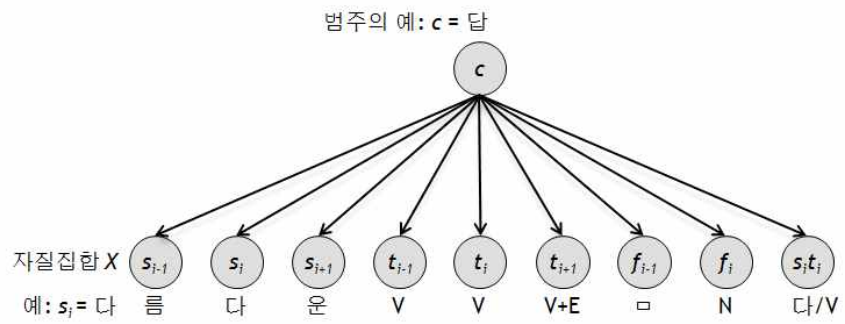
$$X = \{x_j | j = 1,...,9\}$$
$$= \{s_{i-1}, s_i, s_{i+1}, t_{i-1}, t_i, t_{i+1}, f_{i-1}, f_i, f_{i+1}, s_i t_i\} \qquad (2)$$

여기서 s_i는 처리대상이 되는 음절이고, s_{i-1}과 s_{i+1}은 각각 s_i의 앞뒤의 음절(syllable)을 의미하고, t_i는 s_i의 음절품사(syllable tag)이고 t_{i-1}과 t_{i+1}은 각각 s_i의 앞뒤의 음절품사를 의미하며, f_i는 s_i의 받침(final consonant)이고 f_{i-1}과 f_{i+1}은 각각 s_i의 앞뒤의 받침을 의미한다. <표 2>는 어절 ‘아름다운ㅗ’에 대한 자질집합을 표현한 것이다.

〈표 2〉 형태소 복원을 위한 자질집합 X

자질 / 음절	음절정보			품사정보			받침정보		음절품사	범주
	s_{i-1}	s_i	s_{i+1}	t_{i-1}	t_i	t_{i+1}	f_{i-1}	f_i	$s_i t_i$	c
아	$	아	름	$	V	V	$	N	아/V	아
름	아	름	다	V	V	V	N	ㅁ	름/V	름
다	름	다	운	V	V	V+E	ㅁ	N	다/V	답
운	다	운	ㅗ	V	V+E	SP	N	ㄴ	운/V+E	ㄴ
ㅗ	운	ㅗ	추	V+E	SP	N	SP	SP	ㅗ/SP	SP

<표 2>에서 ‘ㅗ’은 공백문자를 나타내며 ‘SP’는 공백문자의 음절품사를 나타낸다. 또한 ‘V+E’은 <표 1>에서 설명한 것과 같이 복합품사이다. NB 모델의 학습은 모든 음절에 대해서 수행하지만 실제 실행할 때는 복합품사를 가지는 음절과 그 앞 음절에만 적용된다[5]. 이렇게 함으로써 형태소 복원을 실행하는 속도는 전체 속도에 거의 영향을 주지 않는다.



(그림 1) 형태소 복원을 위한 NB 모델

4) 참고로 이 모델은 NB 모델 이외 어떠한 모델을 사용해도 무관하나 범주의 수가 많아서 학습 속도를 고려해서 본 논문에서는 NB 모델을 사용한다.

5) 참고로 뒤 음절에서는 형태 변이가 발생되지 않아서 적용하지 않았다.

〈표 3〉 ETRI 구문구조 말뭉치의 통계치

구분	전체	학습	실험
문장 수	101,602	91,658	9,944
어절 수	2,200,914	1,996,613	204,301
형태소 수	4,375,332	3,940,706	434,626

〈표 4〉 KTS@KMU의 전체 시스템의 성능

시스템	실험 말뭉치		입력형태	시스템 결과		Acc	P	R	F_1
	종류	전체 대상수		전체 대상수	정답수				
음절품사 부착기	음절품사 부착	900,153	음절단위 문장	861,011	838,045	97.3%	·	·	·
형태소 복원기	형태소 복원 (전체음절)	900,153	음절품사 부착 결과	900,573	899,074	·	99.8%	99.9%	99.9%
	형태소 복원 (적용음절)	52,418		52,838	51,339	·	97.2%	98.0%	97.6%
품사 부착기	품사 부착	434,626	형태소 복원 결과	435,509	424,064	·	97.4%	97.6%	97.5%

5. 실험 및 평가

본 절에서는 본 논문에서 제안한 형태소 복원 모델을 이용한 한국어 품사 부착 시스템(KTS@KMU)[11]의 성능을 평가하고 제안된 형태소 복원을 위한 NB 모델을 분석한다. 먼저 형태소 복원 모델을 포함하는 한국어 품사 부착 시스템을 간단히 기술하고자 한다. KTS@KMU은 3단계(음절품사 부착기, 형태소 복원기, 품사 부착기)로 구성되었다. 음절품사 부착기는 CRF++[13]을 이용해서 주어진 각 음절에 대한 품사를 부착하며 기본적인 방법은 [5-6]과 같다. 형태소 복원기는 4장에서 제안한 NB 모델을 이용해서 형태소를 복원한다. 품사 부착기는 [5-6]과 같으며 연속되는 음절이 같은 음절품사를 가지면 하나의 형태소로 결합하여 최종 품사 부착 결과를 생성한다. 실험 및 평가를 위해서 사용된 말뭉치는 한국전자통신연구원 ETRI 구문구조 말뭉치[12]이며 그 구성은 <표 3>과 같으며 학습말뭉치와 실험말뭉치는 약 9:1의 비율로 나누었다.

성능 평가 척도로는 정보검색 분야에서 널리 사용되는 정확도(accuracy, Acc), 정확률(precision, P), 재현율(recall, R), F_1-점수(F_1-score, F_1)를 이용한다[14].

<표 4>는 KTS@KMU의 전체 성능평가 결과이며 음절 부착기의 결과가 형태소 복원기의 입력이 되고, 또한 형태소 복원기의 출력이 품사 부착기의 입력이 되는 실질적인 시스템이다. 음절품사 부착기의 성능은 입력 음절의 수와 출력 범주의 수가 같기 때문에 정확도로 평가되어야 하며 97.3%의 정확도를 보였다. 4장에서 설명했듯이 형태소 복원기는 모든 음절에 적용하는 것이 아니라 복합품사를 가지는 음절과 그 앞 음절에만 적용한다. 실험에서 그 대상 수는 52,838개였으며 이들을 대상으로만 성능을 평가하면 F_1점수는 각 97.6%이다[6]. 형태소 복원기의 성능이 음절품사 부착기보다 성능이 좋은 이유는 많은 형태소가 복원되어 실험 말뭉치의 음절과 같게 되었고 그 결과 성능이 좋아질 수 있었다. 형태소 단위의 품사 부착기의 F_1점수는 97.5%로 확률기반의 한국어 품사 부착 시스템[3-4]과 성능을 비교하면 큰 차이가 없다.

(그림 2)는 전체 오류(1,449개)를 유형에 따라 분석한 결과이며 그 유형은 말뭉치 일관성 오류, 복합태그 오류, 형태소 복원기 오류로 분류된다. 많은 경우(약 67%)가 말뭉치의 일관성 오류이다. 예를 들면 정답말뭉치에는 '했'이 '하+었' 혹은 '하+았'으로 분석되어 있어 형태소 복원기는 '하+었'으로 복원되나 '하+았'이 정답인 경우가 있었고 또한 그 반대의 경우도 있었다. 이는 음절품사 부착 말뭉치를 구축할 때 선어말어미 '았'과 '었'을 정규화할 필요가 있었다. 또 다른 예로는 '슬기로운'에서 '로'는 '롭'으로 정확히 복원되지만 '운'은 'ㄴ'이 아닌 '은'으로 복원되어 오류로 평가되었다. 이는 매개모음 '으'에 대한 정규화가 필요한 경우이다. 다음으로 많이 차지하는(약 18%) 오류가 복합태그 오류이다. 이는 음절품사 부착기에서 일반태그를 복합태그로 잘못 인식하여 그 결과가 형태소 복원기에 영향을 준 경우이며 오류 전파에 해당하며 이 결과는 음절품사 부착기에 완벽하지 않는 한 피할 수 없는 오류 중 하나이다. 마지막으로 형태소 복원기 자체 오류로서 복원하지 말아야하는 '로'를 '롭'으로 복원하는 등의 경우이다. 이 경우가 전체 오류 중에 약 15%를 차지했다.

6) 형태소 복원기는 입력 음절이 복원되면서 그대로 형태만 변하는 경우도 있고 두 개의 형태소로 나누어지는 경우도 있기 때문에 입력 음절 수와 출력 범주의 수가 다르다. 이와 같이 일반적으로 입력과 출력의 수가 다르거나 실험 말뭉치과 시스템의 결과의 수가 다를 경우에는 정확도로 평가할 수 없고 정확률과 재현율로 평가되어야 한다.

(그림 2) 형태소 복원 오류의 유형 및 그 비율

이와 같은 오류는 학습말뭉치 정제와 음절품사 부착기 및 형태소 복원기의 성능 개선을 통해서 개선될 수 있을 것이다. 특히 형태소 복원기의 자질에 대해서 좀 더 심도 있는 연구가 필요할 것으로 생각된다.

6. 결론 및 향후 연구 과제

일반적으로 한국어 품사 부착 시스템은 한국어 형태소 분석기를 전처리기로 사용한다. 그러나 한국어 형태소 분석기는 매우 복잡한 구조를 가지고 있으며 복잡한 지식과 방대한 사전 정보가 요구된다. 이러한 문제점을 해결하기 위해 형태소 분석 없이 기계학습 방법을 이용한 한국어 품사 부착을 시도하였다[5-6]. 이러한 시스템도 여전히 복합명사 미분석 문제와 형태소 복원 불완전성 문제를 가지고 있었다. 본 논문에서는 전자의 문제를 음절품사 부착 말뭉치의 부호화 방법으로 해결하였으며 후자의 문제는 NB 모델을 이용해서 해결하였다. 본 논문에서 제안된 기계학습 방법을 이용한 형태소 복원 모델을 제안함으로써 한국어 품사 부착 시스템의 모든 단계가 기계학습 방법으로 구현될 수 있게 되었다. 본 논문에서 제안된 한국어 품사 부착 시스템(KTS@KMU)은 F_1 점수가 97.5%로 형태소 분석기를 포함하는 다른 한국어 품사 부착 시스템과 비슷한 성능을 보였다.

그러나 본 논문에서 제안된 모델은 형태소 복원 범주가 복원 음절 그 자체이므로 범주의 수가 많아서 CRF와 같은 모델에 쉽게 적용할 수 없다. 따라서 형태소 복원 모델의 범주를 불규칙 현상의 유형으로 한다면 좀 더 좋은 결과를 기대할 수 있을 것이다. 그러나 이렇게 하기 위해서는 음절품사 부착 말뭉치를 구축하기 위해서 전문가의 노력이 필요한 실정이다. 또한 음절품사 부착 모델에서 범주에 '+'가 덧붙으므로 범주의 수가 2배로 늘어났다. 이도 개선할 여지가 있을 것으로 생각된다.

참 고 문 헌

[1] A. R. Martinez, "Part-of-Speech tagging", WIREs Computational Statistics, Vol.4, pp.107-113, 2012.

[2] P. J. Antony and K. P. Soman, "Parts Of Speech Tagging for Indian Languages: A Literature Survey", International Journal of Computer Applications, Vol.34, No.8. pp.22-29, 2011.

[3] 김재훈, "가중치망 모델을 이용한 한국어 품사 태깅", 한국정보과학회논문지, 제 25권 제 6호, pp.951-959, 1998.

[4] 임희석, 김진동, 임해창, "통계 정보와 언어 지식의 보완적 특성을 고려한 혼합형 품사 태깅", 정보과학회논문지B, 제 25권 제 11호, pp.1705-1715, 1998.

[5] 김재훈, 이공주, "사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정", 정보처리학회논문지B, 제 10권 제 1호, pp.47-56, 2003.

[6] 심광섭, "형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅", 인지과학, 제 22권 제 3호, pp.327-345, 2011.

[7] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[8] X.-H. Phan, CRFTagger: CRF English POS Tagger, http://crftagger.sourceforge.net/, 2006.

[9] I. G. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: An open-source CRF reference string parsing package", Proceedings of the Language Resources and Evaluation Conference (LREC 08), pp.661-667, 2008.

[10] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning", Proceedings of the 3rd Workshop on Very Large Corpora (ACL 1995), pp.82 - 94, 1995.

[11] 전길호, 기계학습을 이용한 음절기반 품사 부착, 한국해양대학교 대학원, 컴퓨터공학과, 석사학위 논문, 2012.

[12] 김재훈 외, 구문구조 부착 말뭉치 구축, 모비코앤시스메타㈜, 최종보고서, 2005.

[13] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[14] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2007.

김 재 훈

e-mail : jhoon@hhu.ac.kr

1986년 계명대학교 전자계산학과(학사)

1988년 한국과학기술원 전산학과 (공학석사)

1996년 한국과학기술원 전산학과 (공학박사)

1988년~1997년 한국전자통신연구원 선임연구원

1997년~현 재 한국해양대학교 IT공학부 교수

2001년~2002년 Information Sciences Institute in University of Southern California 방문연구원

2007년~2008년 Beckman Institute in University of Illinois at Urbana-Champaign 방문연구원

관심분야 : 자연언어처리, 한국어정보처리, 정보검색, 정보추출

전 길 호

e-mail : jhoon@hhu.ac.kr

2010년 한국해양대학교 IT공학부(학사)

2012년 한국해양대학교 컴퓨터공학과 공학석사

관심분야 : 자연언어처리, 한국어정보처리, 정보검색