(제 10회 한글 및 한국어 정보처리 학술대회)

어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델†

김진동 이상주 임해창

고려대학교 컴퓨터학과
서울시 성북구 안암동 1가, 우: 136-071
jin@nlp.korea.ac.kr zoo@nlp.korea.ac.kr rim@nlp.korea.ac.kr

Morpheme-Unit POS Tagging Model Considering Eojeol-Spacing

Jin-Dong Kim Sang-Zoo Lee Hae-Chang Rim
Department of Computer Science & Engineering
Korea University

요약

한국어 품사 태깅 모델은 어절 단위 모델과 형태소 단위 모델로 나눌 수 있다. 이들 중 형태소 단위 모델은 자료 부족 문제가 별로 심각하지 않고 비교적 풍부한 태깅 결과를 내어 준다는 점에서 선호되나 어절 단위로 띄어쓰기를 하는 한국어의 특성을 제대로 반영하지 못한다는 단점이 있다. 이에 본 논문에서는 한국어의 어절 띄어쓰기 정보를 활용하는 형태소 단위 품사 태깅 모델을 제안한다. 어절 띄어쓰기 정보는 복잡도가 매우 작기 때문에 모델 구축에 드는 추가 비용이 그리 크지 않다. 그럼에도 불구하고 실험 결과는 어절 띄어쓰기 정보가 한국어 품사 태깅에 유용한 정보임을 보여준다.

1. 서론

한국어 품사 태깅 모델은 품사 태깅의 단위에 따라 어절 단위 품사 태깅 모델과 형태소 단위 품사 태깅 모델로 구분할 수 있다.[1]

어절 단위의 품사 태깅 모델은 품사 집합을 어절 단위로 정의하고, 이를 통하여 품사의 전이와 어휘의 발생을 관측하고 품사를 태깅하는 모델이다. 이 모델은 영어권에서 사용되는 품사 태깅 모델을 한국어에 적용시키기 위해서 한국어의 어절을 영어의 단어 개념으로 사상(mapping)시킨 것이다. 그러나 영어의 단어와는 달리 한국어의 어절은 그 형태가 매우 다양하게 발생하기 때문에 어절 단위 통계 정보를 획득할 때 자료 부족 문제가 심각하게 발생한다. 따라서 어절 단위 품사 태깅 모델에서는 단순화된 품사 집합을 사용할 수밖에 없으며 이로 인해 품사 태깅으로 얻어지는 정보가 제한된다.

형태소 단위 품사 태깅 모델은 품사 집합을 형태소 단위로 정의하고, 이를 통하여 품사의 전이와 어휘의 발생을 관측하고 품사를 태깅하는 모델로서, 한국어의 형태소를 영어의 단어 개념으로 사상시킨 것이라고 할 수 있다. 이 모델은 형태소 단위의 통계 정보를 요구하기 때문에 어절 단위의 통계 정보를 요구하는 어절 단위 품사 태깅 모델에 비해서 자료 부족 문제가 덜 심각하다. 따라서 세분화된 품사 집합을 사용하는 것이 가능해지고, 형태소 단위의 품사 정보를 얻을 수 있는 등 어절 단위 품사 태깅 모델의 단점이 많이 해소된다. 그러나 형태소 단위 품사 태깅 모델은 한국어의 중요한 특징 중의 하나인 어절 정보를 이용하지 못한다는 단점이 있다.

최근에는 어절 단위 품사 태깅 모델과 형태소 단위 품사 태깅 모델의 장점을 모두 수용하기 위한 복합 모델에 대한 연구도 보고되고 있다. 이러한 복합 모델들은 형태소 단위 정보와 어절 단위 정보를 적절히 결합시켜서 이를 품사 태깅 정보로 활용하려는 시도라고 볼 수 있다. 그러나 기존의 복합 모델들에서는 형태소 단위 정보와 결합되는 어절 단위 정보들의 복잡도가 크기 때문에 이에 대한 정보를 획득할 때 자료 부족 문제가 심각하게 발생하게 된다.

본 논문에서 제안하는 태깅 모델은 형태소 단위의 품사 태깅 모델을 기본으로 하되 여기에 어절 띄어쓰기 정보를 포함시킴으로써 어절에 관한 정보를 반영하고 하는 것이다. 어절 띄어쓰기 정보는 어절에 관련된 정보이지만 정보의 복잡도가 매우 작기 때문에 모델 구축에 드는 추가 비용이 그리 크지 않다. 그럼에도 불구하고 실험 결과는 어절 띄어쓰기 정보가 한국어 품사 태깅에 유용

† 본 논문은 1997년도 한국과학재단 특정기초연구과제 연구비 지원에 의한 것입니다.

1) 본 논문에서는 통계 기반 품사 태깅 모델만을 논의의 대상으로 한다.

한 정보를 제공함을 보여준다.

2. 기존 연구

2.1 영어 품사 태깅 모델

영어를 위한 품사 태깅 모델로는 1983년 LOB 코퍼스를 태깅하기 위한 CLAWS 시스템[1] 이후 품사 2-gram이나 3-gram 정보를 이용하는 통계 모델이 주로 사용되어져 왔다. 이들 통계 모델들에서는 품사 태깅의 문제를 "길이가 $n(\geq 1)$인 단어열(문장) $w_{1,n}=w_1, w_2, \cdots, w_n$이 주어졌을 때, 이에 대응되는 가장 확률이 높은 태그열 $c_{1,n}=c_1, c_2, \cdots, c_n$을 구하는 것"으로 식 (1)과 같이 정의한다. 여기에서 w_i는 문장에서 i번째에 나타나는 단어를 나타내며 c_i는 i번째 단어에 할당되는 품사를 의미한다.

$$T(w_{1,n}) \stackrel{\text{def}}{=} \underset{c_{1,n}}{argmax}\, P(c_{1,n}|w_{1,n}) \quad (1)$$

식 (1)은 문장 단위의 통계 정보를 필요로 하는 매개변수(parameter) $P(c_{1,n}|w_{1,n})$를 가진다. 그러나 문장 단위의 통계 정보를 획득하는 것은 거의 불가능한 일이기 때문에 식 (1)에 몇몇 가정을 적용시켜 통계 정보 획득이 가능한 형태로 바꾸어야만 한다. 이 때 어떠한 가정을 적용시키느냐에 따라 다양한 통계 모델들이 만들어 질 수 있다.

이들 통계 모델 중에서 가장 많이 사용되는 것은 은닉 마르코프 모델(Hidden Markov Model)에 기반한 태깅 모델로서 식 (2)와 같이 표현된다.

$$T(w_{1,n}) \cong \underset{c_{1,n}}{argmax} \prod_{i=1}^{n} P(c_i|c_{i-1})P(w_i|c_i) \quad (2)$$

이 모델은 식 (1)로부터 유도되는 이론적이 바탕이 견고하고, 태깅된 혹은 태깅되지 않은 말뭉치로부터 모델을 자동으로 학습시킬 수 있으며 정확도가 높다는 장점 때문에 현재 영어의 품사 태깅에 가장 널리 사용되고 있으며 사실상의 표준 모델이라고 할 수 있다[2].

2.2 어절 단위 한국어 품사 태깅 모델

어절 단위 한국어 품사 태깅 모델은 한국어의 띄어쓰기 단위인 어절을 영어의 단어 개념으로 사상시켜서 식 (2)와 같은 영어 품사 태깅 모델을 별다른 수정 없이 한국어 품사 태깅에 사용하고자 하는 것이다. 이 모델에서는 품사 태그 대신에 어절 태그를 사용하게 되며 품사 태깅의 문제는 "길이 $n(\geq 1)$인 어절열(문장) $e_{1,n}=e_1, e_2, \cdots, e_n$이 주어졌을 때, 이에 대응되는 가장 확률이 높은 어절 태그[2]열 $\bar{c}_{1,n}=\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_n$을 구하는 것"으로 식 (3)과 같이 정의된다.

$$T(e_{1,n}) \stackrel{\text{def}}{=} \underset{\bar{c}_{1,n}}{argmax}\, P(\bar{c}_{1,n}|e_{1,n}) \quad (3)$$

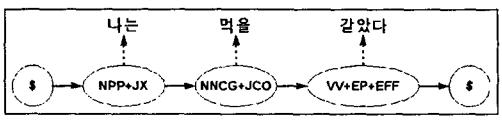
$$\cong \underset{\bar{c}_{1,n}}{argmax} \prod_{i=1}^{n} P(\bar{c}_i|\bar{c}_{i-1})P(e_i|\bar{c}_i) \quad (4)$$

식 (4)는 식 (3)에 마르코프 가정(Markov assumption)을 적용하여 유도한 것으로서 은닉 마르코프 모델에 기반한 어절 단위 태깅 모델이라고 할 수 있다. 이 모델은 한국어의 어절을 영어의 단어와 같이 취급했기 때문에 어절 단위의 통계 정보를 필요로 한다. 그러나 영어의 단어에 비해 한국어의 어절은 훨씬 복잡한 구조를 가지고 있기 때문에 어절 태그 집합이 매우 크며, 이에 대한 통계 정보를 획득할 때 자료 부족 문제 또한 매우 심각하게 나타난다. 따라서 어절 단위 품사 태깅 모델을 사용할 때는 태그 집합을 단순화시켜야 한다는 제약이 따른다. 실제로, 초기의 어절 단위 품사 태깅 시스템인 이운재(1993)의 시스템에서는 400개의 어절 태그를 사용하였으며 이는 형태소 품사를 17개로 분류한 것에 해당한다[3]. 또 다른 어절 단위 품사 태깅 시스템인 이상주(1994)의 시스템에서는 291개의 어절 태그를 사용하였으며 이는 형태소 품사를 14개로 분류한 것에 해당한다[4]. 그러나 이렇게 방대한 태그 집합을 사용함에도 불구하고 어절 단위 품사 태깅 시스템이 제공하는 정보는 다음과 같은 이유 때문에 충분하다고 볼 수 없다.

- 어절 태그 집합의 크기 제약 때문에 품사를 충분히 분류할 수 없다.
- 어절의 형태소 분리에 대한 모호성이 해결되지 않는다.

어절의 형태소 분리에 대한 모호성이 해결되지 않는다는 것은, 예를 들어 어절 '가는'의 어절 태그가 '동사+어미'인 것으로 결정된 다음에도 이에 대해 '가[동사]+는[어미]' 또는 '갈[동사]+는[어미]' 등의 두 가지 이상의 해석이 가능한 경우를 말하는 것이다.

[그림 1]은 식 (4)에서 사용되는 어절 단위 HMM에 의한 한국어 문장의 발생을 보여준다. 이 그림에서 문장에 나타나는 어절 '나는', '먹을', '갈았다'는 각각 어절 태그 'NPP+JX', 'NNCG+JCO', 'VV+EP+EFF'로 태깅되었음을 볼 수 있다.



[그림 1] 어절 단위 HMM에 의한 한국어 문장의 발생 예

2.3 형태소 단위 한국어 품사 태깅 모델

어절 단위 품사 태깅 모델에서는 어절을 태깅의 단위로 하였기 때문에 많은 문제점이 발생하였다. 이에 반해 형태소 단위 품사 태깅 모델은

2) 본 논문에서 $\bar{c}$는 어절 태그를, c는 형태소 태그(품사)를 나타낸다.

형태소를 태깅의 단위로 하여 자료 부족 문제를 완화시키고 태깅의 결과로 어절의 형태소 분리 위치에 대한 모호성도 해소하고자 하는 모델이다.

형태소 단위 품사 태깅 모델에서는 품사 태깅의 문제를 "길이가 $n(\geq 1)$인 어절열(문장) $e_{1,n}=e_1e_2\cdots e_n$이 주어졌을 때, 이에 대응되는 가장 확률이 높은 형태소열 $m_{1,x}=m_1m_2\cdots m_x$과 형태소 태그열 $c_{1,x}=c_1c_2\cdots c_x$을 구하는 것"으로 식 (5)와 같이 정의한다. 이 때, 형태소열 $m_{1,x}$는 문장 $e_{1,n}$을 구성하는 형태소의 열을 의미하며, 형태소 태그열 $c_{1,x}$는 $m_{1,x}$에 대응하는 형태소 태그의 열을 의미한다. x는 문장 $e_{1,n}$을 구성하는 형태소의 개수를 의미하며, 문장 내 어절들의 형태소 분석 결과에 따라 가변적이다.

$$T(e_{1,n}) \stackrel{\text{def}}{=} \underset{m_{1,x},c_{1,x}}{argmax}\, P(m_{1,x},c_{1,x}|e_{1,n}) \quad (5)$$

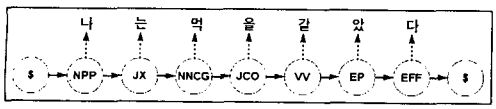
$$\cong \underset{m_{1,x},c_{1,x}}{argmax} \prod_{i=0}^{x} P(c_i|c_{i-1})P(m_i|c_i) \quad (6)$$

식 (5)에서 품사 태깅의 결과로 형태소 태그열($c_{1,x}$) 뿐만 아니라 형태소열($m_{1,x}$)까지 구해줌으로써 어절 단위 품사 태깅의 문제점인 어절의 형태소 분리 문제를 해결할 수 있다. 식 (6)은 은닉 마르코프 모델에 기반한 형태소 단위 품사 태깅 모델이다.

이 모델은 한국어의 형태소를 영어의 단어로 사상시킨 것으로서, 형태소 단위의 통계 정보를 필요로 한다. 이 때, 형태소 태그 집합은 같은 정보량을 가지는 어절 태그 집합에 비해 크기가 훨씬 작으므로 어절 단위 태깅 모델과는 달리 태그 집합을 단순화할 필요가 없다. 실제로, 형태소 단위 품사 태깅 시스템인 임철수(1994)의 시스템에서는 79개의 형태소 태그를 사용하였으며[5], 이상호(1995)의 시스템에서는 54개의 형태소 태그를 사용하였다[6]. 형태소 단위 품사 태깅 시스템은 이렇게 상대적으로 작은 태그 집합으로도 다음과 같은 이유 때문에 어절 단위 품사 태깅 시스템보다 훨씬 많은 정보를 제공한다고 할 수 있다.

- 태그 집합을 단순화하지 않아도 되기 때문에 품사가 충분히 분류될 수 있다.
- 어절의 형태소 분리에 대한 모호성이 해소된다.

[그림 2]는 식 (6)에서 사용되는 형태소 단위 HMM에 의한 한국어 문장의 발생을 보여준다. 이 그림에서 문장에 나타나는 형태소 '나', '는', '먹', '을', '갈', '았', '다'는 각각 형태소 태그 'NPP', 'JX', 'NNCG', 'JCO', 'VV', 'EP', 'EFF'로 태깅 되었음을 볼 수 있다.



[그림 2] 형태소 단위 HMM에 의한 한국어 문장의 발생 예

그러나 그림에서 알 수 있듯이 한 어절의 내부에서 발생하는 품사 전이 'NNP→JX'와 어절과 어절 사이에서 발생하는 품사 전이 'JX→NNCG'가 동등하게 취급되고 있어, 형태소 단위 품사 태깅 모델에 한국어의 어절 구조가 제대로 반영되지 않았음을 볼 수 있다.

2.4 복합 모델

어절 단위 품사 태깅 모델과 형태소 단위 품사 태깅 모델이 각각 장·단점을 가지고 있기 때문에 이들의 장점을 모두 수용하고 단점을 보완하기 위한 복합적인 성격을 띠는 품사 태깅 모델에 대한 연구도 보고된 바 있다.

이러한 연구들에서는 대체로 어절 단위 품사 태깅 모델을 기본으로 하되 자료 부족 문제를 해소하기 위해 일부 어절 단위 통계 정보를 형태소 단위 통계 정보로 근사시키거나[7], 형태소 단위 품사 태깅 모델을 기본으로 하면서 어절에 관한 정보를 일부 활용하는 방법을 사용한다[8, 9].

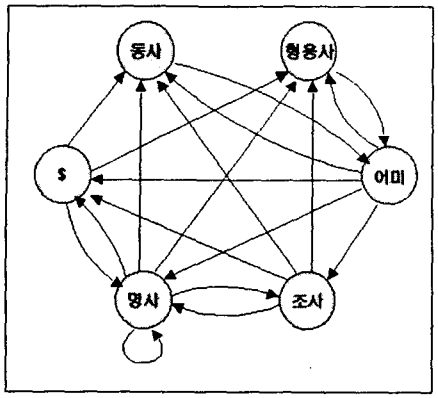
이하규(1997)의 품사 태깅 모델은 전자에 해당하는 모델이다. 이 모델은 기본적으로 어절 단위 품사 태깅을 수행하지만 어절 태그 대신에 어절의 형태소 분석열과 품사열이 결합된 토큰(token)을 태깅의 단위로 사용하기 때문에 형태소 단위의 태깅 결과를 얻을 수 있다. 또한 문맥을 관측할 때 어말-어두의 공기 정보만을 사용함으로써 문맥 정보 획득시의 자료 부족 문제를 감소시켰다. 그러나 어휘의 발생을 어절 단위로 관측하기 때문에 어휘 발생 정보 획득시의 자료 부족 문제는 매우 심각하다. 따라서 단순화된 품사 집합(13개)을 사용해야 하는 어절 단위 품사 태깅 모델의 단점을 극복하지는 못했다[7].

신중호 등(1994)은 어절 정보를 반영한 형태소 단위 품사 태깅 모델을 제안하였다. 이 모델은 기본적으로 형태소 단위 품사 태깅을 수행하지만 문맥을 관측할 때 어절 태그와 형태소 태그를 함께 고려하도록 해서 문맥에 의한 변별력을 높였다. 그러나 어절 정보 획득시에 나타나는 자료 부족 문제에 대한 해결 방법은 제시하지 않았다[8].

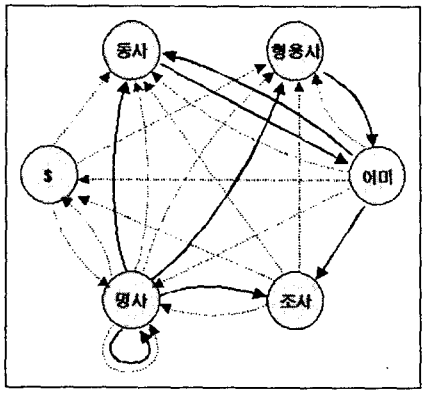
김진동 등(1997)은 어절 구조를 반영한 형태소 단위 품사 태깅 모델(Twoply HMM)을 제안하였다. 이 모델은 기본적으로 형태소 단위 품사 태깅을 수행하면서 어절에 관한 정보를 구조적으로 품사 태깅에 반영하려고 시도하였다. 이 때 모델 구축에 요구되는 정보는 모두 형태소 단위 정보이기 때문에 자료 부족 문제가 그리 심각하게 나타나지 않으면서 어절 구조를 품사 태깅에 반영할 수 있었다. 하지만 모델의 구조가 매우 복잡하기 때문에 모델 구축 비용이 다소 크고 태깅 알고리즘이 복잡하다는 단점이 있었다[9].

3. 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델

본 논문에서는 어절 띄어쓰기를 하는 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델을 제안한다. 이 모델은 기본적으로 식 (5)와 같은 정의에서부터 유도되는 형태소 단위 모델이기 때문에 자료 부족 문제가 경미하고 풍부한 태깅 결



[그림 3] 형태소 단위 품사 전이 모델



[그림 4] 띄어쓰기를 고려한
형태소 단위 품사 전이 모델

과를 제공하는 등 형태소 단위 모델의 장점을 그대로 가지고 있다. 여기에 한국어의 어절 띄어쓰기 정보를 추가하게 되는데, 어절 띄어쓰기 정보는 정보의 복잡도가 매우 낮기 때문에 전체 모델의 복잡도를 거의 상승시키지 않는다.

먼저 식 (6)으로 표현되는 기존의 형태소 단위 품사 태깅 모델의 품사 전이 모델을 [그림 3]을 통해 살펴보자. 편의상 한국어에 5개의 품사만 존재하는 것으로 가정하였으며, 품사는 정점(node)으로, 각 품사간의 전이는 방향 간선(directed edge)으로 나타내었다. 그림에서 알 수 있듯이 기존의 형태소 단위 품사 전이 모델에서는 어절 내부에서 일어나는 품사 전이(예: 명사→조사, 동사→어미 등)와 어절과 어절간에 일어나는 품사 전이(예:조사→동사, 어미→명사 등)가 구분이 되지 않는다. 그렇기 때문에 품사 태깅 시에 한국어의 어절 띄어쓰기 정보를 반영할 수 없다.

한국어의 경우 전이가 일어나는 두 개의 품사가 정해지면 어절내 전이인지 어절간 전이인지도 결정적으로(deterministically) 정해지는 경우가 많지만, 어절내 전이와 어절간 전이가 선택적으로(selectively) 일어나는 경우도 존재한다. [그림 3]의 품사 전이 모델에서 일어날 수 있는 품사간 전이를 어절간 전이와 어절내 전이로 구분하여 [표 1]에 나타내었다[3]. 이 표에서 보여지듯이 몇몇 품사쌍은 어절내 전이와 어절간 전이가 선택적으로 발생하게 되는데, 이 때 어절내 전이인지, 어절간 전이인지에 따라 다른 확률 분포가 적용되는 것이 바람직하다. 그러나, 기존의 형태소 단위 품사 전이 모델에서는 이를 반영할 수 없다.

[표 1] 품사간에 존재하는 전이의 종류
(↗: 어절내 전이, ⇗:어절간 전이)

후 / 전	$	명사	동사	형용사	조사	어미
$		⇗	⇗	⇗		
명사	⇗	↗ ⇗	↗ ⇗	↗ ⇗	↗	
동사						↗
형용사						↗
조사	⇗	⇗	⇗	⇗		
어미	⇗	⇗	↗ ⇗	↗ ⇗	↗	

이와 같은 점을 보완하기 위해 본 논문에서는 어절내 품사 전이와 어절간 품사 전이를 구분해서 표현하는 [그림 4]와 같은 품사 전이 모델을 제안한다. [그림 4]에서 어절내 품사 전이는 실선 간선(solid edge)으로, 어절간 품사 전이는 점선 간선(dotted edge)으로 표현되었다. 어절내 품사 전이와 어절간 품사 전이가 선택적으로 존재하는 품사쌍 사이에는 실선 간선과 점선 간선이 모두 존재함을 볼 수 있다. [그림 3]과 같은 모델을 [그림 4]와 같이 수정할 경우 이론적으로 모델 내에 존재하는 전이의 수가 두 배로 증가하기 때문에 모델의 복잡도 또한 2배로 증가하게 된다. 그러나 실제적으로는 어절내 전이와 어절간 전이 중 오직 한가지 경우로만 발생하는 전이(예: 명사→조사, 어미→명사 등)가 다수 존재하기 때문에 모델의 복잡도 증가는 그리 크지 않다[4]

띄어쓰기를 고려한 품사 전이 모델에 기반한 형태소 단위 품사 태깅 모델은 식 (7)과 같다.

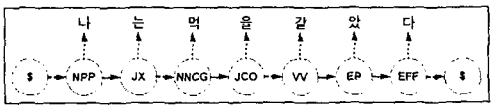
$$T(e_{1,n}) \cong \underset{m_{1,n},c_{1,n}}{argmax} \prod_{i=0}^{n} P(c_i|c_{i-1},k)P(m_i|c_i) \quad (7)$$

$$\begin{cases} k=0 & : \text{if 어절내 전이일 때} \\ k=1 & : \text{if 어절간 전이일 때} \end{cases}$$

3) 이와 같은 품사 전이 모델은 형태소 분석 방법에 따라 다소 차이가 있을 수 있다.

4) [그림 3]과 [그림 4]에 나타낸 예의 경우 전이의 수는 19개에서 23개로 증가하였다. 즉 모델의 복잡도는 $\frac{23}{19}\times 100 \approx 121(\%)$로 증가하였다.

식 (7)에서 k는 품사 전이가 어절내 전이인지 어절간 전이인지를 나타내는 이진(binary) 변수이며, 그 값에 따라 다른 확률 분포가 사용된다.



[그림 5] 띄어쓰기를 고려한 형태소 단위 HMM에 의한 한국어 문장의 발생 예

[그림 5]는 식 (7)에서 사용되는 형태소 단위 HMM에 의한 한국어 문장의 발생을 보여준다. 이 그림에서 문장에 나타나는 형태소 '나', '는', '먹', '을', '갈', '았', '다'가 각각 형태소 태그 'NPP', 'JX', 'NNCG', 'JCO', 'VV', 'EP', 'EFF'로 태깅 된 것은 [그림 2]와 같다. 그러나 어절내 품사 전이(실선으로 표현, 예: 'NNP→JX')와 어절간 품사 전이(점선으로 표현, 예: 'JX→NNCG')가 다르게 취급(다른 확률분포가 적용)되고 있어, 형태소 단위 품사 태깅 모델에 한국어의 어절 띄어쓰기가 반영되었음을 알 수 있다.

4. 실험 및 평가

본 논문에서 제안된 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델의 타당성을 검증하기 위하여, 제안된 모델을 기반으로 한 품사 태깅 시스템을 구현하여 태깅의 정확도를 측정하는 실험을 하였다. 또한, 기존의 한국어 품사 태깅 모델 중 가장 일반적인 모델이라고 할 수 있는 형태소 단위 품사 태깅 모델에 기반한 품사 태깅 시스템[6]도 함께 구현하여 동일한 환경에서 동일한 실험을 수행함으로써 제안된 모델의 상대적인 평가가 이루어지도록 하였다.

구현된 품사 태깅 시스템을 학습시키고 시스템의 품사 태깅 정확도를 측정하기 위하여 태깅된 말뭉치가 사용되었다. 이 때 사용된 말뭉치는 167,115개의 어절로 이루어졌으며 태깅에 사용된 품사 집합은 42개의 형태소 태그와 18개의 기호 태그를 포함한다. 원시 말뭉치의 형태소 분석을 위해 사용된 형태소 사전은 160,118개의 표제어를 포함한다. [표 2]에 실험에 사용된 말뭉치의 대략적인 통계가 요약되어 있다.

[표 2] 실험에 사용된 말뭉치에 대한 통계

문장 개수	어절 개수	형태소 개수	중의성 없는 어절	어절 평균 중의성
15,211개	167,115개	359,031개	44,647개(27%)	3.4개

품사 태깅 실험은 내부 실험과 외부 실험의 두 가지 유형으로 수행되었다. 내부 실험은 실험에 사용된 말뭉치 전체를 사용하여 태깅 시스템을 학습시키고, 학습된 태깅 시스템을 사용하여 말뭉치 전체를 재태깅하였을 때, 원래 태그와의 일치도를 측정하는 실험이다. 이 때 정확도는 어절 단위로 측정하였으며, 어절의 형태소 분리와 분리된 각 형태소의 품사 결정이 모두 정확하게 이루어진 경우에만 정확하게 태깅된 것으로 인정하였다. [표 3]은 내부 실험으로 측정된 품사 태깅 시스템의 정확도를 보여준다. 제안된 모델을 사용한 경우 기존의 형태소 단위 품사 태깅 모델을 사용할 때에 비해 0.08%의 정확도 향상을 보이고 있다. 이러한 수치는 정규 검정(normal test)을 사용할 때 90%의 신뢰수준에서 통계적으로 의미있는 향상이라고 할 수 있다.

[표 3] 품사 태깅 시스템에 대한 내부 실험 결과 (정확도 %)

형태소 단위 HMM (이상호95)	띄어쓰기를 고려한 형태소 단위 HMM
95.72	95.80

외부 실험은 학습에 사용되지 않은 문서에 대한 품사 태깅 시스템의 태깅 정확도를 측정하기 위해 수행되었다. 본 실험에 사용된 말뭉치의 크기가 매우 작기 때문에, 실험 결과의 신뢰도를 높이기 위하여 순환적 실험 방법을 사용하였다. 순환적 실험의 첫 번째 실험에서는 전체 말뭉치의 첫 번째 문장으로부터 시작하여 매 10번째에 오는 문장들을 모두 제외시킨 나머지 문장들(전체 말뭉치의 90%분량)을 이용하여 품사 태깅 시스템을 학습시키고, 제외된 문장들(학습에 포함되지 않은 부분, 전체 말뭉치의 10% 분량)에 대해 품사 태깅 실험을 수행하였다. 두 번째 실험에서는 전체 말뭉치의 두 번째 문장으로부터 시작하여 매 10번째에 오는 문장들을 모두 제외시킨 나머지 문장들을 이용하여 품사 태깅 시스템을 학습시키고, 제외된 문장들에 대해 품사 태깅 실험을 수행하였다. 이런 식으로 순환적 실험에서는 총 10번의 실험을 수행하게 된다. [표 4]는 순환적 외부 실험으로 측정된 품사 태깅 시스템의 정확도를 보여준다. 실험적 수치에서 보여지는 정확도의 향상 정도는 비록 통계적으로 의미있는 향상이라고는 할 수 없지만, 제안된 모델을 사용하는 경우가 기존의 형태소 단위 품사 태깅 모델을 사용할 때에 비해 항상 좋은 정확도를 보임을 확인할 수 있다.

[표 4] 품사 태깅 시스템에 대한 외부 실험 결과 (정확도 %)

언어모델 / 실험No.	형태소 단위 HMM (이상호95)	띄어쓰기를 고려한 형태소 단위 HMM
1	94.04	95.00
2	95.00	95.05
3	95.09	95.15
4	94.99	95.03
5	95.39	95.43
6	95.21	95.23
7	95.01	95.02
8	94.83	94.83
9	95.02	95.03
10	95.06	95.09
평균	95.05	95.09

품사 태깅 시스템의 오류를 분석해 본 결과 오류의 상당 부분은 품사의 세분류에 따른 오류(예: 보통 명사, 동작성 보통 명사, 상태성 보통 명사의 혼돈에 의해 발생하는 오류)들이었다. 이러한 오류들은 품사 태깅 시스템을 사용하는 응용 분야에 따라 허용되는 경우가 많다. 따라서 구현된 품사 태깅 시스템의 유용성을 실용적 관점에서 평가하는데 참고가 될 수 있도록, 순수 한국어 어절에 관련된 44개의 태그를 [표 5]와 같이 12개의 태그로 단순화하였을 때의 품사 태깅 정확도를 측정하는 실험을 수행하여 그 결과를 [표 6]에 나타내었다. 단순화된 품사 집합을 사용할 경우 제안된 모델과 기존의 형태소 단위 품사 태깅 모델 모두 상당히 높은 정확도를 나타내고 있다.

[표 5] 단순화된 품사와 세분된 품사

단순화된 품사	세분된 품사
명사	보통명사, 동작성 보통명사, 상태성 보통명사, 의존명사, 단위성 의존명사, 대명사, 명사 추정
대명사	인칭대명사, 지시대명사
수사	수사, 수관형사
접두사	명사 접두사, 수사 접두사
접미사	명사형 접미사, 관형사형 접미사, 부사형 접미사, 동사형 접미사, 형용사형 접미사
조사	주격 조사, 보격 조사, 목적격 조사, 관형격 조사, 부사격 조사, 호격 조사, 접속 조사, 보조사
서술격조사	서술격 조사
관형사	일반 관형사, 지시 관형사
부사	성상 부사, 서술 부사, 지시 부사, 접속 부사
감탄사	감탄사
용언	동사, 형용사, 보조 용언, 용언 추정
어미	어말 어미, 연결 어미, 부사형 어미, 명사형 어미, 관형사형 어미, 선어말 어미

[표 6] 단순화된 품사 집합에 의한 실험 결과 (정확도 %)

언어모델 / 실험방법	형태소 단위 HMM (이상호95)	띄어쓰기를 고려한 형태소 단위 HMM
내부 실험	98.55	98.58
외부 실험	98.31	98.32

이상의 실험 결과를 종합해 볼 때 제안된 모델이 기존의 형태소 단위 품사 태깅 모델보다 한국어 품사 태깅에 보다 더 적합한 모델이며 띄어쓰기 정보가 한국어 품사 태깅에 도움을 주는 유용한 정보임을 알 수 있다.

5. 결론

본 논문에서는 어절 띄어쓰기를 하는 한국어의 특성을 반영하는 형태소 단위의 품사 태깅 모델을 제안하였다. 제안된 모델은 매우 단순한 구조를 가지고 있기 때문에 구현이 용이하며, 어절 정보의 반영을 위해 정보의 복잡도가 낮은 띄어쓰기 정보만을 활용하기 때문에 모델 구축시에 자료 부족 문제가 별로 심각하게 발생하지 않는다.

제안된 모델의 타당성을 검증하기 위해 제안된 모델에 기반한 품사 태깅 시스템을 구현하여 태깅의 정확도를 측정하는 실험을 수행하였으며, 이때 기존의 형태소 단위 품사 태깅 모델에 기반한 품사 태깅 시스템도 함께 구현하여 동일한 환경에서 실험함으로써 제안된 모델의 상대적인 평가가 가능하도록 하였다. 실험결과, 제안된 모델이 한국어 품사 태깅에 적합한 모델이며 한국어의 경우 어절 띄어쓰기 정보가 한국어 품사 태깅에 유용한 정보임을 알 수 있었다.

추후에는 본 연구에서 제안된 모델을 어절 띄어쓰기 오류 교정에 적용할 계획을 가지고 있으며, 이와 같은 연구를 통해 띄어쓰기 오류에 견고한 한국어 품사 태깅 모델을 개발하고자 한다.

참고문헌

[1] Roger Garside, Geoffrey Leech, Geoffrey Sampson, The Computational Analysis of *English*, *Longman*, *New York*, USA, Chapter 3, 1987

[2] Eugene Charniak, Curtis Hendrickson, Neil Jacobson, Mike Perkowitz, "Equations for part-of-speech tagging," Proc. of the 11th National Conference on Artificial Intelligence(AAAI), pp.784-789, 1993.

[3] 이운재, 한국어 문서 태깅 시스템의 설계 및 구현, 한국과학기술원 전산학과 석사학위논문, 1993.

[4] 이상주, 은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅, 고려대학교 전산과학과 석사학위논문, 1994.

[5] 임철수, HMM을 이용한 한국어 품사 태깅 시스템 구현, 한국과학기술원 전산학과 석사학위논문, 1994.

[6] 이상호, 미등록어를 고려한 한국어 품사태깅 시스템 구현, 한국과학기술원 전산과학과 석사학위논문, 1995.

[7] 이하규, "어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소," 정보과학회 논문지(B), 제24권, 제1호, pp.82-89, 1997

[8] 신중호, 한영석, 박영찬, 최기선, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅," 제 6회 한글 및 한국어 정보처리 학회 논문집, pp.389-394, 1994.

[9] 김진동, 임희석, 임해창, Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델, 정보과학회논문지(B), 24권 12호, pp.1502-1512, 1997