품사 분포와 Bidirectional LSTM CRFs를 이용한 음절 단위 형태소 분석기

김혜민○, 윤정민, 안재현, 배경만, 고영중
동아대학교 컴퓨터공학과
{kimhyemin369, yjungmin2, anjaehyun17, kmbae0722, youngjoong}@gmail.com

Syllable-based Korean POS Tagging using POS Distribution and Bidirectional LSTM CRFs

Hyemin Kim○, Jungmin Yoon, Jaehyun An, Kyoungman Bae, Youngjoong Ko
DongA University, Department of Computer Engineering

요 약

형태소 분석기는 많은 자연어 처리 영역에서 필수적인 언어 도구로 활용되기 때문에 형태소에 대한 품사를 결정하는 것은 매우 중요하다. 최근 음절 기반으로 형태소의 품사를 태깅하는 방법에 대한 연구들이 많이 진행되고 있다. 음절 단위 형태소 분석은 음절 단위로 분리된 형태소에 대해서 기계학습을 이용하여 분리된 음절 단위로 품사를 태깅하는 단계를 가진다. 본 논문에서는 기존의 CRF를 이용한 음절 단위 품사 태깅 방법을 개선하기 위해 bi-LSTM-CRFs를 이용한 방법을 제안한다. 또한, bi-LSTM-CRFs의 입력을 음절의 품사 분포 벡터를 이용해 확장함으로써 음절 단위 품사 태깅의 성능을 향상 시켰다.

주제어: 음절 단위 형태소 분석, Bidirectional LSTM, CRF

1. 서론

한국어 형태소 분석의 부정확한 결과는 구문 분석, 의미역 부착, 기계 번역 등에 치명적인 영향을 미칠 수 있으므로 정확한 분석이 중요하다[1]. 형태소 분석은 일반적으로 형태소 분석과 품사 태깅 두 가지로 나뉜다. 형태소 분석이란 가장 작은 의미를 가진 형태소와 품사 쌍 후보를 생성하는 것이다. 그리고 품사 태깅이란 형태소 분석에서 나온 후보들에서 각 어절의 뜻과 문맥을 고려하여 가장 알맞은 형태소와 품사 쌍을 결정하는 것이다[1,2].

기존의 형태소 단위로 한국어 어절을 분석하기 위해서는 형태소 복원과 동시에 형태소 단위의 분리 과정, 형태소에 대한 품사 결정 과정이 함께 필요하다. 각 과정에서 형태적 중의성 및 품사적 중의성이 발생하므로 이를 처리하기 위한 과정이 비교적 복잡하다[3-5]. 최근에는 이를 해결하기 위해 음절 단위 품사 태깅에 대한 연구가 늘어나고 있다. 음절 단위 품사 태깅은 어절 단위로 품사 태깅 할 때 보다 자료 부족 문제가 줄어들고, 띄어쓰기 등의 기능과 결합이 가능하며, 다른 언어 이식과 이전 연구보다도 우수한 성능을 보인다[2,6].

음절 단위 형태소 분석은 입력된 문장을 음절단위로 나누고, CRF와 같은 기계학습 기반 분류기를 이용해 음절 단위로 형태소 시작과 이어지는 형태소를 나타내는 B,I 태그가 포함된 품사 레이블을 결정한다. 그리고 한국어는 교착어로 다양한 음운 현상이 발생하기 때문에 효과적인 음절 단위 형태소 분석을 위해서 기분석 사전을 이용한다[7]. 기분석 사전은 형태소 분석이 된 어절들을 특정한 기준을 통해 미리 만들어 놓고 품사 태깅 시 이용하는 것이다. 또한, 불규칙 용언을 해결하기 위해 원형복원 사전을 추가적으로 이용한다[2]. 원형복원 사전은 복합 형태소를 대상으로 간단한 규칙을 통하여 복합태그를 부착하는데 사용한다.

음절 단위 형태소 분석을 위해서는 앞서 언급한 것과 같이 순차적 레이블링을 처리할 수 있는 기계학습 기반의 분류기가 필요하며, Structural SVM와 CRF를 이용한 음절단위 형태소 분석 연구가 있다. 본 논문에서는 최근 순차 레이블이 많은 영역에서 좋은 성능을 보이고 있는 Bidirectional Long Short Term Memory CRFs (이하 bi-LSTM-CRFs)을 이용한 음절 단위 형태소 분석 방법을 제안한다.

Structural SVM와 CRF는 한 음절에 대해 레이블을 결정하기 위해 다양한 자질을 사용해야한다. 특히 현재 음절의 앞과 뒤에 존재하는 음절 또는 어절에 대한 정보를 자질로 활용하는 것이 중요하다. 반면, bi-LSTM-CRFs은 forward 단계에서 현재 입력에 대한 상태층의 정보가 뒤의 상태에 영향을 주며, backward 단계에서 뒤에 상태가 앞의 상태에 영향을 주어 학습이 되기 때문에 다른 순차 레이블링을 위한 기계학습과 달리 작은 수의 자질만으로도 좋은 결과를 얻을 수 있다.

bi-LSTM-CRFs은 음절단위로 입력이 결정되며, 각 음절에 대한 음절 표상을 나타내는 벡터를 구성하여 입력으로 사용된다. 본 논문에서는 이를 위해 대용량 원시 말뭉치를 기반으로 음절 표상을 나타내는 64차원의 벡터를 생성하여 입력으로 사용하였다. bi-LSTM-CRFs의 성능을

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2015R1D1A1A01056907)

향상시키기 위해 본 논문에서는 각 음절이 전체 학습 말뭉치에서 출현한 품사 태그의 분포를 벡터로 구성하여 bi-LSTM-CRFs의 입력을 확장하였다. 각 음절은 학습 말뭉치에서 여러 형태소에 포함이 될 수 있으며, 여러 형태소는 다양한 품사를 가질 수 있다. 학습 말뭉치에서 음절이 포함된 형태소의 품사 빈도수를 계산한 후 softmax를 통해서 확률을 각 차원의 값으로 사용하였다. 본 논문에서 제안하는 음절 단위 형태소 분석기는 bi-LSTM-CRFs으로 음절 단위 품사 태그를 결정한 후 기존의 방법과 같이 기분석 사전과 원형 복원 사전을 적용하여 최종적으로 형태소 분석된 결과를 보여준다.

음절의 품사 분포 벡터를 이용한 bi-LSTM-CRFs을 음절 단위 형태소 분석기에 효과적으로 적용한 결과 기존의 CRF 기반의 음절 단위 형태소 분석기에 비해 3.01%가 향상된 97.09%의 성능을 보였다. 이를 통해 제안한 음절의 품사 분포와 bi-LSTM-CRFs이 음절 단위 형태소 분석기에 효과적이라는 것을 확인 할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 설명하고, 3장에서는 제안하는 음절의 품사 분포와 bi-LSTM-CRFs을 이용한 제안 방법을 설명한다. 4장에서는 제안한 방법을 실험을 통해 평가하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

한국어 형태소 분석에 관한 다양한 연구가 진행되었다[2-6]. [2]는 순차적 레이블링 기반 음절 단위 품사 태깅 방법의 전처리 단계로 품사 태깅말뭉치와 국어사전으로부터 구축된 복합명사 기분석사전과 약 1천만 어절의 세종 품사 태깅말뭉치로부터 자동 추출된 어절 사전을 적용함으로써 품사 태깅 성능을 개선시키는 방법을 제안하였다. [3]은 품사 태깅 말뭉치로부터 자동 추출된 음절 n-gram 정보, 음절 복원 정보, 태그 바이그램 정보를 이용하는 음절 단위의 한국어 형태소 분석 모델을 제안하였다. 제안한 모델에서는 원형 복원을 하기 전에 주어진 어절의 각 음절에 대한 품사 태깅을 먼저 하는데, 이는 원형 복원을 먼저 하는 기존 확률 모델에 비하여 형태소 분석 과정이 훨씬 효율적이고 간결한 방법을 제안하였다. [4]는 한국어 형태소 분석을 위한 3단계 확률 모델을 제안하였다. 이 모델은 분석 단계를 형태소 복원, 분리, 태깅의 3단계로 나누어 독립된 모듈로 처리함으로써 기존의 2단계 확률 모델보다 처리 복잡도를 줄였다. 또한, 음절 대신 자소 단위의 처리를 하고, 형태소 전이 확률을 이용하여 형태소 분리를 함으로써 다양한 품사 태깅 원칙을 학습할 수 있는 방법을 제안하였다. [6]은 음절 단위의 한국어 품사 태깅에서 문제점으로 지적된 바 있는 원형 복원 문제에 대한 새로운 해결 방안을 제시하였다. 이 방법에서는 품사 태깅 말뭉치로부터 자동 생성된 음절 복원 사전을 이용하여 원형 복원을 수행한다. 이 과정에서 복잡한 한국어 형태론적 처리를 하지 않아도 되므로 음절 태깅 후 형태소를 구성하는 과정이 매우 단순화된다는 장점을 가지고 있는 방법을 제안하였다.

3. 제안 방법

최근의 형태소 분석을 위해서 기계학습을 이용한 음절 기반의 형태소 품사 태깅에 대한 연구가 이루어지고 있다. 품사 태깅의 순서는 그림 1과 같다.

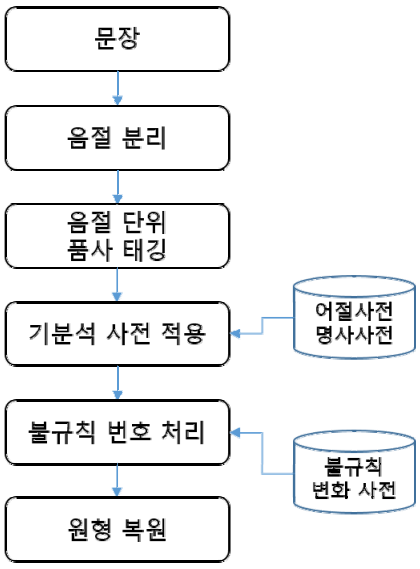

그림 1. 음절 기반 형태소 품사 태깅 구성도

먼저 들어온 문장을 음절 단위로 나눈다. 분리된 음절은 CRF와 같은 기계학습 기반의 분류기를 통해 음절이 포함된 형태소의 품사 태그를 할당받는다. 음절 단위로 품사 태그가 결정된 결과에 대해 기분석 사전을 통해 학습 말뭉치에서 중의성이 없는 변환을 처리하여 오류를 줄인다. 마지막으로 음절 단위의 결과를 원형복원을 통해 형태소 단위의 결과로 만들어 준다. 원형복원 단계에서는 불규칙 변환이 존재하는 경우 불규칙 변환 사전을 통해 이를 보정하여 변환한다.

3.1 CRF를 이용한 음절 단위 품사 태깅

음절 기반의 형태소 품사 태깅을 위해서 분리된 음절에 대해 품사 태그를 결정하는 것이 중요하다. 예를 들어, 문장에서 “세계적인”이라는 어절은 음절 단위로 분리된 후 각 음절별로 아래와 같이 품사 태그가 할당된다.

표 1. 음절 단위 품사 태그 부착의 예

세 B-NNG
계 I-NNG
적 I-NNG
인 B-VCPDIC

B-NNG는 품사가 NNG인 형태소의 시작 음절을 나타내며, I-NNG는 품사가 NNG인 형태소의 이어진 음절을 나타낸다. 음절 단위로 품사 태그 부착하기 위해서 CRF와 같

은 기계학습 기반 분류기를 학습하여야 한다. 이를 위해서 각 음절에 대해 아래 표 2와 같은 자질들을 사용한다[2].

표 2. 음절 단위 품사 태그 부착을 위한 CRF 학습에 사용된 자질 유형

Feature	Explanation
Unigram 음절	x_{t-1}, x_t, x_{t+1}
Bigram 음절	$x_{t-2}x_{t-1}, x_{t-1}x_t, x_tx_{t+1}, x_{t+1}x_{t+2}$
Trigram 음절	$x_{t-2}x_{t-1}x_t, x_{t-1}x_tx_{t+1}, x_tx_{t+1}x_{t+2}$
Unigram 어절	t_{t-1}, t_t, t_{t+1}
Bigram 어절	$t_{t-2}t_{t-1}, t_{t-1}t_t, t_tt_{t+1}, t_{t+1}t_{t+2}$
Trigram 어절	$t_{t-2}t_{t-1}t_t, t_{t-1}t_tt_{t+1}, t_tt_{t+1}t_{t+2}$

만일 형태소가 불규칙 형태소였으면 음절 단위로 부착되는 품사 태그의 뒤에 “DIC” 태그를 부착하여 불규칙 변환 사전을 적용해야 할 형태소임을 나타낸다. “DIC” 태그가 부착된 음절은 원형복원 단계에서 불규칙 변환 사전을 이용하여 불규칙에 대한 문제를 해결한다. 예를 들어, “세계적인”의 경우 “세계적/NNG+이/VCP+ㄴ/ETM의상/NNG”과 같이 품사 태깅이 이루어 지며, 여기서 “인”은 CRF 학습시에 “인 B-VCPDIC”의 태그를 가지고 학습에 사용된다.

본 논문에서는 CRF 학습을 위해 표 2에서 언급한 음절 단위 자질과 어절 단위 자질을 사용한다. 어절 단위 자질은 효과적인 사용을 위해 전체 말뭉치에서 유일한 어절들을 추출한 후 각 어절 별로 ID를 할당하고, 할당된 ID를 자질로 표현하여 사용한다. 추가적으로 문장의 시작은 :S를 나타내고, trigram 어절에서 :S 이전의 어절 위치 역시 :S를 사용하였다. 또한, 문장의 끝은 :O를 이용해 표현하였다. 예를 들어, 문장의 시작과 끝인 경우 아래 표 3과 같이 자질이 생성된다.

표 3. CRF 학습을 위한 어절 자질의 예

문장	단어	생성된 자질						
		음절	-2	-1	0	1	2	태그
시작	목욕	목	:S	:S	36	37	38	B-NNG
		욕	:S	:S	36	37	38	I-NNG
끝	품목.	품	164	98	165	:O	:O	B-NNG
		목	164	98	165	:O	:O	I-NNG
		.	164	98	165	:O	:O	B-SF

3.2 Bidirectional LSTM-CRFs를 이용한 음절 단위 품사 태깅

CRF는 하나의 음절에 대한 품사 태그를 부착하기 위해 여러 자질을 사용해야 한다. 반면, bi-LSTM-CRFs은 음절에 대한 벡터만을 입력으로 사용하여 좋은 결과를 얻을 수 있다. 아래 그림 2는 음절을 입력으로 하는 bi-LSTM-CRFs의 예이다.

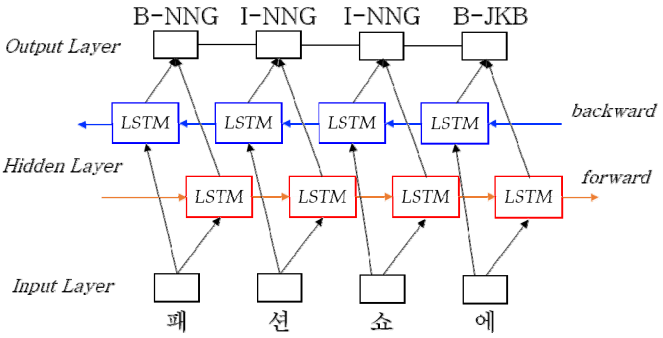

그림 2. 음절을 입력으로 사용한 bi-LSTM-CRFs의 예

bi-LSTM-CRFs은 음절단위로 학습을 하기 위해 “패션쇼에”라는 어절이 들어왔을 때 forward 단계에서 먼저 “패”라는 음절이 입력되고, 다음으로 음절 “션”이 입력이 된다. 음절 “션”이 입력되었을 때 이전 음절의 상태가 현재 음절의 상태에 영향을 주어 현재 음절의 상태는 실제로는 “패션”을 나타내는 상태와 같은 의미를 가지게 된다. 모든 음절이 입력으로 들어갈 때까지 bi-LSTM-CRFs의 forward를 진행한다. backward 단계에서는 forward와 반대로 음절 “에”가 먼저 입력이 되고, 다음으로 음절 “쇼”가 입력이 된다. 어절에 대한 forward와 backward 단계가 진행된 후 두 단계를 결과와 정답과의 비용(cost)을 계산한 후 역전파(back-propagation) 알고리즘을 통해 학습한다. [8]는 태그 사이의 전이 확률을 반영하여 성능을 개선하는 연구를 진행하였으며, 이를 위해 CRF와 같이 forward 알고리즘을 이용하고, 최적의 태그열을 찾기 위해 Viterbi search 알고리즘을 이용하였다. 본 논문에서도 bi-LSTM-CRFs를 사용하여 음절 단위 품사 태깅을 진행한다. bi-LSTM-CRFs은 앞의 음절과 다음 음절의 정보가 반영되어 학습이 되기 때문에 CRF와 달리 음절의 입력만을 사용하기 때문에 효과적이다. 이러한 장점을 가지는 bi-LSTM-CRFs을 이용한 음절 기반의 형태소 품사 태깅 방법을 제안한다.

3.2.1 bi-LSTM-CRFs의 음절 임베딩

bi-LSTM-CRFs을 사용하기 위해서는 입력되는 음절에 대한 벡터가 필요하다. 음절에 대한 N차원의 벡터는 차원별로 가중치를 가지게 된다. 기본적인 가중치로 랜덤하게 실수를 할당할 수 있다. 본 논문에서는 음절에 대한 벡터를 생성하기 위해 대표적인 단어 임베딩(word embedding) 알고리즘인 word2vec를 사용하여 64차원의 음절 단위의 임베딩 벡터를 학습하여 입력 벡터로 사용하였다.

3.2.2 음절의 품사 분포 벡터

본 논문에서는 bi-LSTM-CRFs를 이용한 음절 기반의 형태소 품사 태깅의 성능 향상을 위해 음절의 입력 벡터를 확장한다. 이를 위해, 음절이 학습 말뭉치에서 포함된 형태소의 품사 분포를 벡터로 표현하여 입력 벡터를 확

장한다. 음절은 포함된 형태소에 따라 다른 품사를 가질 수 있다. 예를 들어, 음절 “하”는 학습 말뭉치에서 명사 태그를 가지는 “하늘”의 일부 음절일 수도 있고, 형용사 태그를 가지는 “하얗게”의 일부 음절일 수도 있다. 이러한 음절이 학습 말뭉치에서 출현한 형태소의 품사 분포를 벡터로 표현하여, bi-LSTM-CRFs의 입력으로 사용하였다. 음절의 품사 분포를 나타내는 벡터는 46개의 품사에 B, I 태그와 DIC 태그가 반영된 131개의 차원에 문장의 처음, 끝, 공백을 나타내는 3개의 태그를 추가한 총 134차원으로 표현된다. 각 차원의 값은 음절에 대한 품사 태그가 말뭉치에서 나온 빈도수를 모두 계산한다. 한 음절에 대해 말뭉치에서 출현한 모든 빈도를 계산한 후 softmax를 통해 확률값으로 만들어서 벡터의 값을 결정한다.

표 4. 음절 “랑”에 대한 품사 분포 벡터의 예

	B-NNP	I-NNP	B-JKG	B-NNG	I-NNG	B-VCPDIC	...
빈도수	2	159	0	0	197	0	...
softmax	0.0046	0.3655	0	0	0.4529	0	...

음절 “랑”은 학습 말뭉치에서 총 435번 출현하였으며, B-NNP, I-NNP, I-NNG, B-JKB 등의 품사 태그를 가지고 있다. 생성한 벡터는 음절 임베딩 벡터와 결합하여 최종적으로 198(64+134)차원의 벡터를 생성하여 bi-LSTM-CRFs의 입력으로 사용한다.

3.3 기분석 사전 적용 및 원형복원

기분석 사전은 품사 태깅에 모호성이 존재하지 않는 경우를 사전으로 구축한 것으로 CRF의 음절별 품사 태그 부착 결과와 상관없이 학습 말뭉치에 존재하는 품사 태그로 변환하기 위해 필요하다[2,7]. 기분석 사전은 어절 사전과 명사사전을 사용한다. 예를 들어, “엔터테이너”라는 명사의 경우 말뭉치에서 “NNG”이외의 품사 태그가 부착되는 경우가 존재하지 않기 때문에 이러한 경우는 명사 사전에 포함한다. 표 5는 명사 사전을 적용한 예이다.

표 5. 명사 사전을 적용한 예

명사	결과
엔 B	엔 B B-NNG
터 I	터 I I-NNG
테 I	테 I I-NNG
이 I	이 I I-NNG
너 I	너 I I-NNG

명사사전은 중의적 분석이 되지 않는 명사들로 구축을 하였다. 하지만, 명사일 경우만 뽑으면 짧은 명사일 경우 “은/NNG”과 “은/JX”인 동일한 글자이나 다른 품사로 태깅된다. 이를 해결하기 위해 일정 길이가 넘는 명사와 복합 명사를 전체 데이터에서 169,004개를 구축하였다.

어절사전은 [2]와 동일하게 문맥정보를 고려하지 않은 어절사전1과 문맥정보를 고려하여 모호성을 해결한 어절사전2를 구축하였다. 어절사전2는 어절만 보았을 때 문맥의 중의성이 있는 어절이 문제가 되므로 이를 해결하기 위해 어절사전에 들어갈 해당 어절의 전 어절에 포함된 마지막 품사 태그를 함께 저장을 하여 중의성을 해결한다. 어절사전1은 1,552,635개, 어절사전2는 37,233개를 구축하였다.

표 6. 어절 사전을 적용한 예

사전	어절	결과	이전품사
어절 사전1	가 B 정 I 하 I 고 I	가 B B-VV 정 I I-VV 하 I I-VV 고 I B-EC	-
어절 사전2	적 B 게 I	적 B B-VA 게 I B-EC	JKB

음절 단위로 품사 태그가 부착한 결과를 최종적으로 원형 복원을 통해 형태소 단위로 변환한다. 원형 복원 시에는 앞서 설명한 불규칙 변환이 필요한 경우 불규칙 변환 사전을 이용하여 변환 한다.

표 7. 불규칙 변환 사전 예

음절 및 태그	변환 결과	빈도수
혀 I I-VVDIC	하/VV+여/EC	5
혀 I I-VVDIC	치/VV+어/EC	4
혀 B B-VVDIC	하/VV+여/EF	1

불규칙 변환에서 불규칙 변환 사전에 동일한 변환이 있을 시 가장 높은 빈도의 결과를 선택한다. 불규칙 변환을 적용한 후 최종적으로 동일한 품사 태그를 가지는 형태소들은 결합하여 형태소 품사 태깅을 완료한다. 예를 들어 “생산/NNG+노동자/NNG+들/XSN” 이면 품사 'NNG'가 동일하기 때문에 "생산노동자/NNG+들/XSN"의 형태로 수정한다.

4. 실험 및 성능 평가

4.1 데이터 셋

본 논문에서는 음절 기반의 형태소 품사 태깅을 평가하기 위해서 세종코퍼스를 사용하였다. 최종적으로 CRF 기반의 방법과 제안하는 bi-LSTM-CRFs를 이용한 방법을 비교하기 위해 랜덤하게 50만 어절을 선택한 후 40만 어절을 학습에 사용하였으며, 10만 어절을 테스트에 사용하였다. 모든 모델에 대해서 40만 어절로 학습을 하는 것은 시간이 많이 소모되기 때문에 제안한 방법 중 가장 좋은 성능을 보이는 모델을 판단하기 위해서 랜덤하게 5만 어절을 선택하여 4만어절로 학습을 하고, 1만 어절로

테스트하여 평가를 진행하였다. 평가를 위해서 정확도(accuracy)를 사용하였다.

$$\text{음절단위 정확도} = \frac{\text{분석된 음절의 품사가 맞은 총 음절수}}{\text{분석해야할 모든 음절수}}$$

$$\text{어절단위 정확도} = \frac{\text{한 어절의 형태소들에 대한 분석결과가 모두 맞은 총 어절수}}{\text{분석해야할 모든 어절의 수}}$$

4.2 음절 단위 품사 태깅의 성능 비교

본 논문에서는 CRF 기반의 음절 단위 품사 태깅과 bi-LSTM-CRFs 기반의 음절 단위 품사 태깅에 대한 시스템을 구축하고 평가를 진행하였다. 두 시스템에 대한 평가는 음절 단위 정확도로 결과를 비교하였다.

표 8. CRF를 이용한 음절 단위 품사 태깅 결과(%)

	4만 어절	40만 어절
CRF(음절자질)	84.34	-
CRF(음절+어절자질)	87.35	**85.28**

CRF를 이용한 음절 단위 품사 태깅의 경우 4만 어절을 이용한 실험을 통해 음절 자질과 어절 자질을 사용했을 때 성능이 좋았다. 본 논문에서는 이를 베이스 시스템(baseline)으로 결정하고, 40만 어절을 이용해 실험을 진행하였다.

표 9. bi-LSTM-CRFs를 이용한 음절 단위 품사 태깅 결과(%)

	4만 어절	40만 어절
bi-LSTM-CRFs(랜덤)	89.83	-
bi-LSTM-CRFs (음절임베딩)	83.58	-
bi-LSTM-CRFs (랜덤+품사분포)	90.81	**92.93**

bi-LSTM-CRFs는 입력 벡터가 필요하며, 본 논문에서는 음절의 임베딩을 생성하여 실험을 진행하였다. word2vec를 이용해 생성한 음절 임베딩보다 랜덤하게 가중치를 준 실험이 더 좋은 성능을 보였다. bi-LSTM-CRFs의 입력은 랜덤을 기반한 벡터를 사용하였으며, 여기에 본 논문에서 제안한 품사 분포를 이용해 입력 벡터를 확장하였을 때 제일 좋은 성능을 얻을 수 있었다. 그림 3을 보면 제안하는 bi-LSTM-CRFs를 이용한 음절 단위 품사 태깅이 CRF를 이용한 방법보다 4만 어절로 학습했을 때 3.46%, 40만 어절로 학습 했을 때 7.65%의 향상된 성능을 보이는 것을 볼 수 있다. 따라서, bi-LSTM-CRFs가 음절 단위 품사 태깅에 효과적이고, 제안하는 음절의 품사 분포가 음절 단위 품사 태깅의 성능 향상에 효과적인 것을 확인할 수 있다.

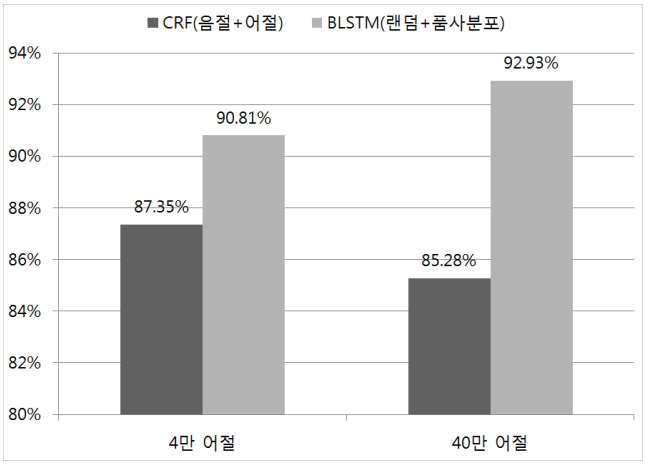

그림 3. CRF와 bi-LSTM-CRFs를 이용한 음절 단위 품사 태깅 결과의 비교(%)

4.3 기분석 사전 및 원형복원

음절을 기반으로 한 형태소 분석은 음절 단위 품사 태깅을 진행한 후 기분석 사전을 적용한 후 원형복원을 통해 형태소 단위의 품사 결과를 얻을 수 있다. 본 논문에서 구축한 기분석 사전인 명사 사전과 어절 사전을 적용하고, 불규칙 변환 사전을 적용한 원형복원 된 결과에 대해 어절 단위 정확도를 이용해 평가하였다. 기분석 사전과 불규칙 변환 사전을 적용한 후 원형복원을 한 결과 제안한 방법이 CRF를 기반한 방법에 비해 3.01% 향상된 성능을 보였다.

표 10. 기분석 사전과 원형복원을 적용한 결과 비교(%)

	40만 어절
CRF(음절+어절자질) +기분석사전+원형복원	94.08
bi-LSTM-CRFs(랜덤+품사분포) +기분석사전+원형복원	**97.09 (+3.01)**

5. 결론

본 논문은 bi-LSTM-CRFs를 이용한 음절 기반의 형태소 품사 태깅 방법을 제안하였다. 음절 단위 품사 태깅을 위해 CRF 대신 bi-LSTM-CRFs를 사용하였으며, 음절의 품사 분포를 벡터로 표현하여, bi-LSTM-CRFs의 입력 벡터를 확장함으로써 음절 단위 품사 태깅의 성능을 개선하였다. bi-LSTM-CRFs를 통해 음절 단위 품사 태깅을 진행한 후 구축된 기분석 사전인 어절사전, 명사 사전을 적용하여 오류를 감소시킨 후 불규칙 변환 사전을 통해 불규칙 변환에 대한 처리를 진행하였다. 마지막으로 원형복원을 통해 음절 단위 결과를 형태소 단위로 복원하여 형태소 품사 태깅을 진행하였다. 본 논문에서 제안한 음절의 품사 분포 벡터를 이용한 bi-LSTM-CRFs 기반의 음절 품사 태깅 방법을 적용하였을 때 CRF 기반의 방법에 비해 7.65% 향상된 92.93%의 음절 단위 품사 태깅 성능

을 보였으며, 기분석 사전, 불규칙 변환 사전을 적용한 후 원형복원 했을 때 CRF보다 3.01% 향상된 97.09%의 성능을 보였다.

향후 bi-LSTM-CRFs의 성능 향상을 위해 확장 가능한 입력 벡터에 대한 연구를 추가적으로 진행할 것이다.

참고문헌

[1] 이승욱, 이도길, 임해창,“형태소 분석 및 품사 부착을 위한 말뭉치 기반 혼합 모형”,*한국컴퓨터정보학회 논문지 제13권 제7호, pp.11-18,* 2008.12

[2] 이충희, 임준호, 임수종, 김현기, “기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅”,*정보과학회논문지 제43권 제3호, pp.362-369,* 2016.3

[3] 심광섭, “품사 태깅 말뭉치에서 추출한 n-gram을 이용한 음절 단위의 한국어 형태소 분석”, *정보과학회논문지 : 소프트웨어 및 응용 제40권 제12호, pp.869-876,* 2013.12

[4] 이재성, “한국어 형태소 분석을 위한 3단계 확률 모델”,*정보과학회논문지 : 소프트웨어 및 응용 제38권 제5호, pp.257-268,* 2011.5

[5] S. S. Kang, "Korean Morphological Analysis using Syllable Information and Multi-word Unit Information", *Seoul National University Computer Engineering Dept. Ph. D. Thesis,* 1993. (in Korean)

[6] 심광섭, "음절 단위의 한국어 품사 태깅에서 원형복원", *정보과학회논문지 : 소프트웨어 및 응용 제40권 제3호, pp.182-189 ,* 2013.3

[7] 곽수정, 김보경, 이재성, “한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법”,*정보처리학회논문지, 제2권 제12호, pp.881-888,* 2013년

[8] 이창기, “Long Short-Term Memory 기반의 Recurrent Neural Network를 이용한 개체명 인식”, *한국정보과학회 2015 한국컴퓨터종합학술대회 논문집,pp. 645-647,* 2015.06,

[9] S. S. Kang, "Korean Morphological Analysis and Information Retrieval", *Hongreung Science Publisher,* 2002. (in Korean)