언어 모델 다중 학습을 이용한 한국어 개체명 인식

김병재, 박찬민, 최윤영, 권명준, 서정연

서강대학교 컴퓨터공학과

Wizard3021@naver.com, cksals302@gmail.com, chldbsdu3773@gmail.com, corundum240@gmail.com,

profseojy@gmail.com

Korean Named Entity Recognition using Joint Learning with Language Model

Byeong-Jae Kim, Chan-min Park, Yoon-Young Choi, Myeong-Joon Kwon, Jeong-Yeon Seo

Sogang University, Dept. of Computer Engineering

요 약

본 논문에서는 개체명 인식과 언어 모델의 다중 학습을 이용한 한국어 개체명 인식 방법을 제안한다. 다중 학습은 1 개의 모델에서 2 개 이상의 작업을 동시에 분석하여 성능 향상을 기대할 수 있는 방법이지만, 이를 적용하기 위해서 말뭉치에 각 작업에 해당하는 태그가 부착되어야 하는 문제가 있다. 본 논문에서는 추가적인 태그 부착 없이 정보를 획득할 수 있는 언어 모델을 개체명 인식 작업과 결합하여 성능 향상을 이루고자 한다. 또한 단순한 형태소 입력의 한계를 극복하기 위해 입력 표상을 자소 및 형태소 품사의 임베딩으로 확장하였다. 기계 학습 방법은 순차적 레이블링에서 높은 성능을 제공하는 Bi-directional LSTM CRF 모델을 사용하였고, 실험 결과 언어 모델이 개체명 인식의 오류를 효과적으로 개선함을 확인하였다.

주제어: 개체명 인식, 다중 학습, 단어 표상, 심층 학습

1. 서론

다중 학습은 서로 다른 여러 작업들을 동시에 학습하는 방법을 말한다. 여러 작업을 동시에 학습하는 동안 작업들 사이의 공통점과 차이점을 활용하여 효과적으로 모델을 학습 할 수 있다. 다중 학습은 개별 모델을 학습하는 것 보다 예측 정확도에 대한 성능 향상을 기대 할 수 있다. 이와 같은 다중 학습은 자연어처리[1]뿐만 아니라 컴퓨터비전[2] 및 음성인식[3]에서도 성공적으로 연구되었다. 하지만 대다수 작업의 경우, 학습되는 데이터에 태그를 부착해야 하는 문제가 발생한다. 이를 해결하기 위해 본 논문에서는 별도의 태그 작업이 필요하지 않은 한국어 언어 모델과 개체명 인식 모델을 다중 학습하는 모델을 제안한다.

개체명 인식은 지명, 인명, 기관명, 날짜, 시간과 같은 고유한 의미를 갖는 단어를 문서에서 추출하고 그 종류를 결정하는 자연어 처리의 한 분야이다. 기존에 기계학습 알고리즘을 사용한 개체명 인식 연구는 사람이 직접 추출한 자질을 입력으로 사용했다. 이러한 방법은 자질을 추출하는데 많은 어려움과 시간이 요구된다. 하지만 최근 개체명 인식을 비롯한 다양한 자연어 처리 분야에 사용되는 심층 학습 모델은 자질 추출 작업 없이 모델을 학습시킬 수 있다는 장점이 있다. 특히 Bi-LSTM-CRF 모델은 많이 쓰이는 심층 학습 모델 중 하나로써 개체명 인식을 비롯한 순차적 레이블링 작업에서 우수한 성능을 보이고 있는 모델로, 입력 단어를 양방향 LSTM 의 입력으로 사용하여 각 입력에 상응하는 출력 계층의 태그간 의존성을 CRF 를 사용하여 모델링한 기법이다.

이와 같은 Bi-LSTM-CRF 모델의 성능은 입력 단어 표상에 의존적이다[4]. 따라서 단어 표상을 확장시켜 개체명 인식 시스템의 성능을 높이기 위한 연구[4,5]가 수행되었다.

본 논문에서는 단어 표상에 사전 학습된 단어 임베딩을 사용하였다. 추가적으로 품사 임베딩, 자소 임베딩 및 개체명 사전을 사용하여 단어 표상을 확장하였다.

또한 본 논문에서는 데이터에 대한 추가적인 레이블링 작업이 필요 없는 언어 모델의 특성을 이용하여 한국어 언어 모델과 한국어 개체명 인식 모델을 동시에 학습하는 다중 학습 모델을 제안한다. 제안하는 다중 학습 모델은 동일한 입력데이터를 효율적으로 학습 할 수 있다. 개체명 인식 모델이 학습되는 동안 입력으로 사용되는 학습 코퍼스에 대해서 동시에 언어 모델을 학습하게 된다. 결과적으로 다중 학습 모델은 사용 가능한 학습 코퍼스를 최대한 활용하는 방향으로 모델을 학습하게 된다. 학습 코퍼스를 최대한 활용함으로써, 은닉 계층을 효율적으로 학습시켜 기존 모델보다 기존 모델에 비해 우수한 성능을 얻을 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구를 소개하고, 3 장에서 제안하는 모델인 Bi-LSTM-CRFs 모델, 단어 표상의 확장 및 멀티 태스크 모델에

대해 소개한다. 4 장에서는 실험 결과를 분석하고, 마지막으로 5 장에서는 결론에 대해서 기술한다.

2. 관련 연구

개체명 인식에서 사용되는 기계학습 알고리즘은 사람이 추출한 자질을 입력으로 받아 최적의 가중치를 학습한다. 대표적인 방법으로 HMM, CRF, Structural SVM[6] 등이 있다. 하지만 최적의 자질 조합을 추출하는 과정에는 많은 연구와 시간이 필요했다. 이와 같은 문제를 해결하기 위해 딥 러닝 기반의 개체명 인식 연구가 많이 진행되고 있다. 특히 순환신경망의 종류 중 하나인 LSTM 모델과 CRF 를 결합한 Bi-LSTM-CRF[7]를 사용한 모델이 좋은 성능을 보였다. 이는 입력 단어의 앞뒤 문맥을 고려한 모델로써, 정방향(forward)과 역방향(backward)을 나타내는 두개의 LSTM 의 은닉 계층을 결합한다. 입력 단어의 Bi-LSTM 출력 결과와 인접 단어의 출력 결과 간의 의존성을 모델링 하기 위해 CRF 를 사용한 모델이다.

LSTM 의 입력으로 사용되는 형태소 단위의 단어 표상을 음절 단위로 세분화 시킨 연구로써 [8]에서는 각 입력 단어 문자열에 K 개의 합성 필터를 적용하여 음절 별 임베딩을 추출했고 이를 각 입력 단어 표상에 확장시켰다. 단어 단위보다 더 작은 문자 단위 입력을 표현하므로 처음 등장한 단어에 대해서도 유연하게 단어 표상을 표현 할 수 있다는 장점이 있다.

[9,10]에서는 학습되는 모델의 은닉 계층을 효율적으로 학습시키기 위해 멀티 태스크 학습을 연구하였다. 멀티 태스크 학습이란 두 가지 이상의 태스크를 파라미터를 공유하며 동시에 학습하는 방법으로 주 태스크에 대해 성능을 높일 수 있다는 장점이 있다. 멀티 태스크 기반 학습은 크게 두가지로 분류되는데 1) Hard parameter sharing 2) Soft parameter sharing 기법이 있다. Hard parameter sharing 은 일반적으로 멀티 태스크 학습에서 사용하는 기법으로써 학습이 진행되는 동안 모든 태스크 사이에 은닉계층을 공유한다. [11]에서는 더 많은 모델을 동시에 학습 할 수록 학습하고자 하는 모델의 과적합을 줄일 수 있음을 보여준다. 반면, Soft parameter sharing 은 Hard parameter sharing 과는 반대로 각 태스크의 모델은 고유의 은닉계층을 소유하는 기법으로 각 모델의 매개변수들이 유사해지도록 하기 위해 모델의 파라미터 사이의 거리를 정규화하는 특징이 있다.두 기법 모두 서로 다른 태스크를 동시에 학습하여 성능을 향상시켰다는 장점이 있다.

본 논문에서는 단어 표상을 표현하기 위해 사전 학습된 단어 임베딩과 품사 임베딩을 사용했고, 추가적으로 자소 단위 임베딩과 개체명 사전 자질을 사용하여 단어 표상을 확장하였다. 또한 모델의 과적합을 줄이고 성능을 높이기 위해 Hard parameter sharing 기법을 적용한 멀티 태스크 학습 모델을 제안한다.

3. 제안 모델

본 논문에서는 한국어 언어 모델 다중 학습을 이용한 한국어 개체명 인식 모델을 제안한다. 제안 하는 모델의 전체 구조는 그림 1 과 같다.

본 논문에서는 개체명 인식 연구에서 좋은 성능을 보이는 Bi-LSTM-CRFs 을 기본 모델을 사용하였다. 입력 단어를 표현하기 위해 사전 학습된 단어, 품사 임베딩을 사용하였고, 자소 임베딩 및 개체명 사전 자질을 사용해 단어 표상을 확장 시켰다. 또한 동일한 입력 데이터를 최대한 활용하기 위해 개체명 인식 모델과 한국어 언어 모델을 동시에 학습시키는 다중 학습 모델을 제안한다.

3.1 Bi-LSTM-CRFs 모델

양방향 LSTM 은 순차적으로 형태소 단위의 단어표상을 입력으로 사용한다. LSTM 의 출력 계층에선 입력 받은 단어 표상의 출력 결과와 인접 출력 결과 간의 의존성을 모델링하기 위해 각 출력 결과를 CRF 에 전달한다. 그림 2 은 기본적인 Bi-LSTM-CRFs 모델의 구조도이다.

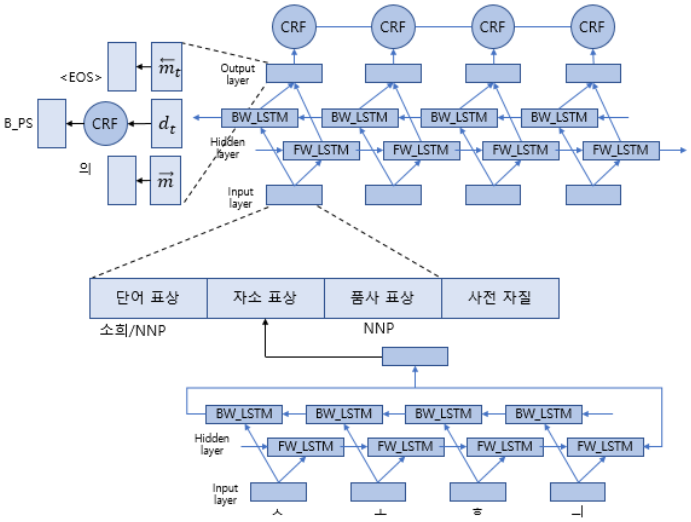

그림 1. 언어 모델 다중 학습을 이용한 개체명 인식 모델의 전체 구성도

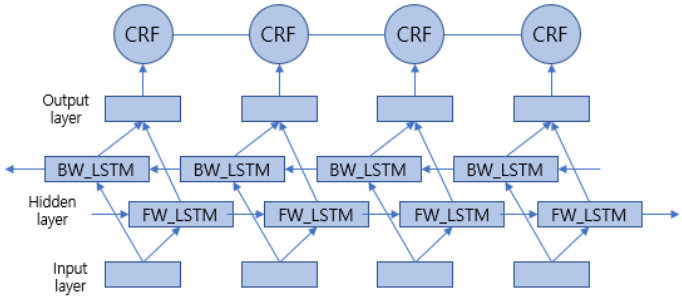

그림 2. Bi-LSTM-CRFs Model

3.2.1 단어 표상 확장

Bi-LSTM 모델의 성능은 단어 표상에 의존적이다[1]. 본 논문에서는 단어 표상을 확장하기 위해 사전 학습된 단어 임베딩, 문자열의 자소 임베딩, 형태소 임베딩, 음절 분포 그리고 개체명 사전 자질을 추가하여 성능을 향상시켰다.

3.2.2 단어 임베딩

제안하는 개체명 인식 모델의 기본 입력은 형태소 단위의 단어이다. 따라서 본 논문에서는 사전 학습된 단어 임베딩을 사용하였다. 또한 한국어의 언어적 특성상 형태소의 품사도 개체명 인식에서 중요하게 사용될 수 있기에 품사 임베딩을 사용하여 단어 표상을 확장했다. 원-핫 인코딩 방식 대신 임베딩 차원을 통해 품사를 표현함으로써, 품사가 가지고 있는 언어적 특성을 잘 표현하였다.

3.2.3 자소 임베딩

단어 표상을 확장하기 위해 입력되는 각 단어를 자소 단위로 분리하였다. 단어마다 구성된 자소의 개수가 다르기 때문에 분리된 자소는 Bi-LSTM 입력으로 사용된다. Bi-LSTM 의 마지막 은닉 계층을 결합하여 하나의 벡터로 변환하고 이를 각 단어를 표현하는데 사용하였다.

3.2.4 음절 분포 임베딩

단어를 구성하고 있는 음절 역시 3.2.3의 자소 임베딩과 같이 Bi-LSTM의 입력으로 사용된다. 음절 분포는 올해 경진대회에서 배포했던 훈련 셋에서 추출하였고 각 태그 별 분포를 나타내는 임베딩을 하였다[4]. Bi-LSTM 의 양방향 최종 출력 값 벡터를 결합하여 이를 전체 임베딩과 결합하여 사용하였다. 자소 임베딩과 달리 학습되지않는다.

3.2.5 개체명 사전

개체명 사전이란 개체명이 될 수 있는 명사들을 사전 형식으로 저장해 놓은 일종의 데이터베이스로, 성능 향상에 중요한 역할을 하는 자질로 사용 된다. 표(1)은 벡터로 표현된 개체명 사전 자질의 예시이다. “소희” 란 입력 단어가 개체명 사전에 등장할 경우, 등장한 태그의 값을 1로 표기하고, 등장하지 않으면 0으로 표기한다. 개체명 사전 벡터의 차원은 개체명 태그 카테고리 수와 같은 5차원이다.

PS	OG	LC	DT	TI
1	0	0	0	0

표 1. “소희”의 개체명 사전 자질 벡터

3.3 멀티 태스크 학습

앞선 3.1의 Bi-LSTM-CRFs 모델은 입력 문장과 이에 대응하는 개체명 태그에만 최적화 된 모델이다. 본 논문에서는 개체명 인식 모델의 학습이 진행되는 동안 입력 문장에 대한 한국어 언어 모델을 동시에 학습하는 다중 학습을 사용한 모델을 제안한다. 그림 1은 제안하는 멀티 태스크 학습 기반 개체명 인식 모델의 전체 구성도이다. 3.1에서 설명한 바와 같 Bi-LSTM-CRFs의 은닉계층인 $\overrightarrow{h_t}$와 $\overleftarrow{h_t}$는 개체명 태그를 예측하기 위해 개체명 출력 계층인 d_t에 서로 합쳐져서 연결된다. 이와 동시에 다중 학습 모델은 양방향 한국어 언어 모델을 학습한다. 예를 들어, “눈물을 훔치며 소희의 마지막을 보러 가야 할 사람은 두환만이 아니었다.” 란 문장이 있고 현재 입력 단어가 “소희” 일 때, $\overrightarrow{h_t}$와 $\overleftarrow{h_t}$는 인접 단어인 “며” 와 “의” 를 예측해야 한다. 따라서 양방향 언어 모델 계층인 $\overrightarrow{m_t}$와 $\overleftarrow{m_t}$는 식(1)과 같이 계산한다.

$$\overrightarrow{m_t} = \tanh\left(\overrightarrow{W_m}\,\overrightarrow{h_t}\right) - (1)$$
$$\overleftarrow{m_t} = \tanh\left(\overleftarrow{W_m}\,\overleftarrow{h_t}\right) - (2)$$

$\overrightarrow{W_m}$과 $\overleftarrow{W_m}$은 학습 가중치를 의미한다. 양방향 언어 모델의 성능은 제안하는 다중 학습 개체명 인식 모델의 주 목적이 아니다. 따라서 본 논문에서는 학습 속도 상향을 위해 m_t의 차원을 h_t에 비해 축소 시켰다. 축소된 차원을 통해 학습 모델은 한국어 언어적 특성에 일반화되어 과적합을 피할 수 다는 장점이 있다[10].

최종적으로 언어 모델 계층인 $\overrightarrow{m_t}$와 $\overleftarrow{m_t}$는 다음 등장 할 단어를 예측하기 위해 식(3,4)와 같이 소프트맥스 함수를 적용해 확률값을 계산한다.

$$P(w_{t+1}|\overrightarrow{m_t}) = softmax\left(\overrightarrow{W_q}\,\overrightarrow{m_t}\right) - (3)$$
$$P(w_{t-1}|\overleftarrow{m_t}) = softmax\left(\overleftarrow{W_q}\,\overleftarrow{m_t}\right) - (4)$$

각 언어 모델의 에러 함수는 순차 입력열 내 단어의 예측 확률에 대한 NLL(Negative Log Likelihood)으로 식(5), (6)와 같이 정의한다.

$$\overrightarrow{E} = -\sum_{t=1}^{T-1} \log\left(P(w_{t+1}|\overrightarrow{m_t})\right) - (5)$$
$$\overleftarrow{E} = -\sum_{t=2}^{T} \log\left(P(w_{t-1}|\overleftarrow{m_t})\right) - (6)$$

LSTM-CRFs 모델의 에러 함수와 각 언어 모델의 에러 함수를 더해 다중 학습을 이용한 개체명 인식 모델의 에러 함수를 정의한다. 최종적으로 본 논문에서 제안하는 개체명 인식 모델의 에러 함수는 다음 식(7)과 같다.

$$\tilde{E} = E + \gamma\left(\overrightarrow{E} + \overleftarrow{E}\right) - (7)$$

가중치 γ는 개체명 인식 모델과 언어 모델 사이의 상대적 중요성을 제어하는 역할을 한다. 본 논문에는 0.1을 사용하였다.

이와 같은 학습 과정을 통해 모델의 은닉 계층은 양방향 한국어 언어 모델을 학습함으로써 한국어의 문법적, 의미적 정보를 학습하게 된다. 은닉계층에 학습된 언어적 자질은 개체명 인식 모델이 개체명 태그를 예측하는데 재사용 된다. 따라서 개체명 인식 모델은 언어의 문법적 특성과 의미적 특성을 활용해 개체명 태그를 예측 할 수 있다.

결과적으로 멀티 태스크 학습 모델은 다음 단어의 등장 확률과 이전 단어의 등장 확률 그리고 현재 입력 단어에 대한 개체명 태그를 예측하는 학습에 최적화된다.

실험을 통해 멀티 태스크 학습을 통한 개체명 인식은 기존 Bi-LSTM-CRFs 모델보다 높은 성능을 보였다.

4. 실험

4.1 실험 환경

제안하는 멀티 태스크 학습을 이용한 Bi-LSTM-CRF 모델의 성능 평가를 위해 사용된 데이터는 2017 년 국어 정보 처리 시스템 경진 대회[12]에서 배포한 데이터를 사용하였다. 학습으로는 3,814 문장, 평가 데이터로 445 문장을 사용하였다. 모든 실험 성능은 F1-measure 평가 방법을 사용하였다. 개체명 인식 모델은 TensorFlow[13]로 구현하여 실험하였다.

4.2 단어 임베딩 실험

사전 학습된 단어 임베딩은 2016 년 국어 정보 처리 시스템 경진대회에서 배포한 데이터를 사용하였다. 임베딩의 차원은 50 차원이며 약 240,000 개의 단어로 구성되어 있다. 단어 임베딩 벡터를 랜덤으로 초기화한 경우보다 사전 학습된 단어 임베딩 벡터를 사용한 경우, F1-score 가 0.68% 높았다.

	prec	recall	F1
RandomVec	85.38	83.94	84.63
Pre-trained	86.02	84.67	85.31

4.3 자소 임베딩 및 품사 임베딩 실험

자소 임베딩의 차원은 50 차원으로 설정하였다. 자소 임베딩은 학습시 랜덤으로 초기화했으며 임베딩에 필요한 LSTM 의 은닉 계층의 차원은 50 차원이다. 자소 임베딩을 학습 할 시 자소 임베딩을 추가하지 않은 경우보다 0.71% 높은 성능을 얻을 수 있었다.

	prec	recall	F1
자소-	86.02	84.67	85.31
자소+	87.01	85.04	86.02

4.4 음절 분포 벡터

음절 임베딩의 차원은 12차원으로 설정하였다. LSTM의 은닉층의 차원도 12차원으로 설정하였다.

	prec	recall	F1
음절-	87.01	85.04	86.02
음절+	86.41	86.57	86.49

4.5 품사 임베딩 실험

품사 임베딩 차원은 16차원으로 설정하였다. 품사 임베딩은 약3.8G의 wiki 데이터를 사용하였다. 학습 모델로는 Word2Vec을 사용하였다.

	prec	recall	F1
품사-	86.41	86.57	86.49
품사+	88.08	86.57	87.75

4.6 개체명 사전 자질 실험

학습에 사용된 개체명 사전은 국어 정보 처리 시스템 경진대회에서 배포한 사전, 위키 코퍼스 인명사전 및 학습데이터에서 추가로 추출한 개체명 사전을 사용하였다. 개체명 사전을 자질로 사용 할 경우, 사용하지 않은 경우보다 1.43%의 향상된 성능을 얻을 수 있다.

	prec	recall	F1
개체명 사전-	88.08	86.57	87.75
개체명 사전+	88.91	89.45	89.18

4.7 멀티 태스크 학습 실험

학습에 사용된 언어 모델 계층 m_t의 차원은 50 차원으로 사용하였다. 동일한 개체명 데이터에 대해 언어 모델을 추가하여 멀티 태스크 학습을 진행 할 경우 개체명 인식 모델만 학습한 경우 보다 성능이 1.49% 증가하였다.

	prec	recall	F1
멀티태스크-	83.98	82.56	83.14
멀티태스크+	85.38	83.94	84.63

5. 결론

본 논문에서는 Bi-LSTM-CRFs를 사용한 개체명 인식 모델을 기반으로 단어 표상을 확장하기 위해 모델 학습 자소 단위 임베딩을 추가하였다. 또한 입력 데이터를 최대한 활용하기 위해 개체명 인식 모델과 한국어 언어 모델을 동시에 학습시키는 멀티 태스크 학습 기법을 적용된 모델을 제안하였다. 실험 결과, 제안하는 멀티 태스크 학습 모델은 멀티 태스크를 사용하지 않은 기본 Bi-LSTM-CRFs 보다 한국어 개체명 인식에서 향상된 성능을 보였다.

참고문헌

[1]Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

[2]Deng, Li, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[3]Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

[4]유홍연, 고영중. "Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장." 정보과학회 논문지, 44.3 (2017.3): 306-313.

[5]나승훈, 민진우. "문자 기반 LSTM CRF 를 이용한 개체명 인식." 한국정보과학회 학술발표논문집, (2016.6): 729-731.

[6]Changki Lee, Junseok Kim, Jeonghee Kim, Hyunki Kim, "Named Entity Recognition using Deep Learning," Korean Institute of Information Scientists and Engineers(KIISE), No. 12, pp. 423-425, 2014.

[7]Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991* (2015).

[8]Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." *arXiv preprint arXiv:1511.08308* (2015).

[9]Zhang, Yu, and Qiang Yang. "An Overview of Multi-Task Learning." *National Science Review* (2017).

[10]Rei, Marek. "Semi-supervised Multitask Learning for Sequence Labeling." *arXiv preprint*

[11]Baxter, Jonathan. "A Bayesian/information theoretic model of learning to learn via multiple task sampling." *Machine learning* 28.1 (1997): 7-39.

[12]https://ithub.korean.go.kr/user/contest/contestIntroLastView.do

[13]https://www.tensorflow.org/