Bidirectional LSTM CRFs를 이용한 한국어 개체명 인식

송치윤, 양성민, 강상우

가천대학교 소프트웨어학과
{a2c222, ysm0622, swkang}@gachon.ac.kr

Named-entity Recognition Using Bidirectional LSTM CRFs

Chi-Yun Song, Sung-Min Yang, Sangwoo Kang
Department of Software Engineering, Gachon University

요 약

개체명 인식은 문서 내에서 고유한 의미를 갖는 인명, 기관명, 지명, 시간, 날짜 등을 추출하여 그 종류를 결정하는 것을 의미한다. Bidirectional LSTM CRFs 모델은 연속성을 갖는 데이터에 가장 적합한 RNN기반의 심층 학습모델로서 개체명 인식 연구에 가장 우수한 성능을 보여준다. 본 논문에서는 한국어 개체명 인식을 위하여 Bidirectional LSTM CRFs 모델을 사용하고, 입력 자질로 단어뿐만 아니라 품사 임베딩 모델과, 개체명 사전을 활용하여 입력 자질을 구성한다. 또한 입력 자질에 대한 벡터의 크기를 최적화 하여 기본 모델보다 성능이 향상되었음을 증명하였다.

주제어 : 개체명 인식, 개체명 사전, Bidirectional LSTM CRFs, 워드 임베딩

1. 서론

개체명(Named-Entity)이란 문장 내에서 인명, 기관명, 지명 등과 같은 고유한 의미가 있는 명사를 의미한다. 개체명 인식은 문장 내에서 개체명을 추출하여 개체명의 카테고리를 파악하는 것이다.

전통적인 개체명 인식 방법은 사전기반, 규칙기반 방법을 사용하였지만 최근 기계 학습 방법을 적용한 연구들이 많이 시도되고 있다. 지도 학습 방법으로는 HMM(Hidden Markov Model)[1], SVM(Support Vector Machine), CRFs(Conditional Random Fields)를 사용하는 방법들이 활발히 연구 및 제안되었고, 최근에는 RNN(Recurrent Neural Network), FFNN(Feed-Forward Neural Network), CNN(Convolutional Neural Network)[3] 등의 심층 신경망을 이용한 방법들이 기존의 지도 학습 방법보다 상대적으로 높은 성능을 보인다. 심층 신경망 모델 중 RNN 모델은 연속성을 갖는 데이터에서 우수한 성능을 보이지만 기울기 손실(vanishing gradient)이라는 문제점을 갖고 있지만 최근 개선된 모델인 LSTM(Long Short-term Memory Network) 모델은 memory cell과 3개의 gate를 통해 RNN의 기울기 손실 문제를 해결한다[2].

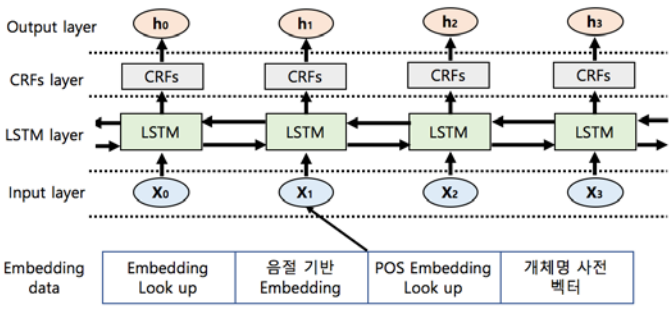
본 논문에서는 기존의 LSTM 모델을 기반으로 데이터를 양방향성(Bi-directional)으로 입력하고, 모델의 출력 값들 사이의 전이 확률을 포함시킴으로써 연속적인 데이터에 적합한 Bi-directional LSTM-CRFs 모델을 적용한다.

개체명 인식의 입력은 기본적으로 형태소 단위를 사용한다. 형태소 단위는 비지도 학습(unsupervised learning)을 통해 사전 학습된 단어 및 품사 임베딩 모델을 사용하며 추가적으로 단어 임베딩 벡터, 개체명 사전 자질 벡터를 통해 입력 단어의 표상을 확장한다.

본 논문의 구성은 다음 장에서 Bidirectional LSTM-CRFs 모델 및 단어 표상 확장 방법에 대해 소개한다. 3장에서는 본 논문에서 제시한 방법에 대한 실험 결과를 평가 및 분석하고, 마지막으로 결론을 기술한다.

2. 제안 방법

본 논문은 [그림 1]과 같이 임베딩 데이터가 Input layer통해 입력 되고 LSTM layer와 CRFs layer를 통해서 예측한 개체명(h)이 출력이 되는 계층 구조를 제안한다.



[그림 1] 제안 모델의 전체 구성도

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학 지원사업의 연구결과로 수행되었음 (2015-0-00932)

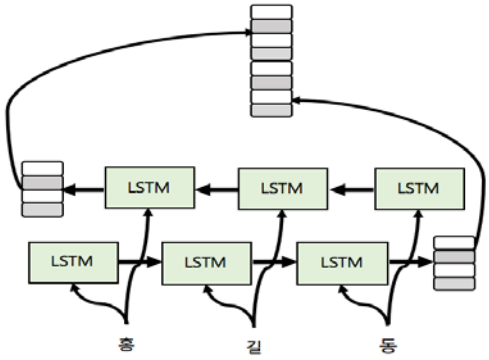
2.1 Bidirectional LSTM-CRFs를 이용한 학습 모델

Bidirectional LSTM-CRFs 모델은 LSTM 모델에서 문자열을 양방향으로 입력 받는다. 각 들은 은닉계층을 통과하고, CRFs를 통해 전이확률을 추가하여 결과 간의 의존성을 고려한다.

2.2 입력 자질

2.2.1 워드 임베딩

제안하는 모델에서는 입력 문장을 벡터로 변환하여 입력 자질로 사용한다. 워드 임베딩 모델을 구축하기 위하여 국립국어원에서 제공된 학습 말뭉치와 세종코퍼스, 위키피디아에서 수집한 말뭉치를 사용하였다. gensim 모델을 통해 워드 임베딩 모델을 사전 학습시켰다. 품사 정보를 반영하기 위해 단어와 품사 태그를 결합한 형태로 적용하였다. 또한, 미등록어 문제를 보완하기 위해서 음절 기반 워드 임베딩을 사용하였다. 이 방법은 미등록어가 입력되었을 때 말뭉치 내에서 유사한 단어들을 기반으로 벡터를 생성해줌으로써 미등록어 문제를 완화할 수 있다. 해당 음절 기반의 워드 임베딩은 [그림 2]과 같이 Bidirectional LSTM을 통해 변환된다.



[그림 2] 입력 벡터 생성 예시

2.2.2 품사 임베딩

품사는 개체명 인식에서 매우 중요한 자질이다. 따라서 품사 정보를 사용하여 입력데이터의 정보량을 확장하였다. 일반적으로 품사 자질을 추가하기 위해서는 원-핫 인코딩(one-hot encoding)을 통해 벡터로 변환하지만 품사들간의 연속적인 의미를 반영하기 위해서 품사 기반의 사전 학습된 벡터를 생성한다. 품사 임베딩 모델은 워드 임베딩 모델과 마찬가지로 국립국어원에서 제공된 학습 말뭉치와 gensim 모델을 사용하여 학습하였다.

2.2.3 개체명 사전

개체명 사전의 정보를 이용하여 입력 단어로부터 자질을 확장한다. 개체명 사전은 2016 ~ 2017년 국어 정보처리 시스템 경진대회에서 제공된 말뭉치와 세종코퍼스, 위키피디아의 데이터에서 추출한 개체명을 활용하였다. 추출된 개체명들은 n-gram 자질로 구성하고 카이제곱 통계량을 사용하여 자질을 선택하였다.

3. 실험 및 평가

제안한 개체명 인식 방법의 성능 평가를 위해 Bidirectional LSTM CRFs를 Tensorflow로 구현하였다. 개체명 인식 평가 데이터로는 2017년 국어 정보 처리 시스템 경진대회에서 배포한 개체명 말뭉치 문장을 사용하였다. 총 4,259 문장 중 3,000 문장을 학습 데이터로, 1,259 문장을 평가 데이터로 사용하였다. 전체적인 실험 성능은 가장 높은 성능을 보인 30 epoch로, 워드 임베딩, 음절 기반 워드 임베딩, 품사 임베딩은 각각 50차원으로 실험하였다. 실험 성능은 F_1-score로 평가하였다.

[표 1] 제안 모델 성능 평가 (%)

Accuracy	RNN	LSTM	Bi-LSTM CRFs
Baseline	73.26	78.16	80.16
+ Word Embedding	75.72	79.30	81.78
+ POS Embedding	76.18	80.07	83.26
+ GAZETTEER	77.27	82.06	**84.62**

[표 1]과 같이 학습 모델로는 RNN, LSTM, Bidirectional LSTM CRFs 모델을 비교하고 입력 자질로는 워드 임베딩, 음절 기반 워드 임베딩, 품사 임베딩, 개체명 사전을 조합하여 구성하였을 때의 성능을 실험하였다. 실험 결과 [표 1]과 같이 Bidirectional LSTM CRFs 모델에 모든 자질을 입력으로 사용한 경우 84.62%로 가장 높은 성능을 나타내었다.

4. 결론

본 논문에서는 한국어 개체명 인식을 위하여 Bidirectional LSTM CRFs 모델을 적용하고 자질로서 단어 임베딩, 품사 임베딩 그리고 사전 정보 자질을 이용하는 방법을 제안하였다.

향후에는 개체명을 인식할 때 개체명의 접두어, 접미어와 같은 세분화된 의미 정보를 사용하여 입력 자질을 일반화 및 최적화 하는 방법을 연구할 계획이다.

참고문헌

[1] N. V. Patil, A. S. Patil, B. V. Pawar, “HMM based Named-entity Recognition for inflectional language,” Computer, Communications and Electronic s, 2017.

[2] G Lample, M Ballesteros, S Subramanian, "Neural Architectures for Named Entity Recognition," NAACL, 2016.

[3] Y Luo, Y Cheng, O Uzuner, P Szolovits, "Segment convolutional neural networks (Seg-CNNs) for

classifying relations in clinical notes," JAMIA, 2017.