Bidirectional LSTM-CRF 모델을 이용한 멘션탐지

박천음○, 이창기
강원대학교
{parkce, leeck}@kangwon.ac.kr

Mention Detection using Bidirectional LSTM-CRF Model

Cheoneum Park○, Changki Lee
Kangwon National University Dept. of Computer Science

요 약

상호참조해결은 특정 개체에 대해 다르게 표현한 단어들을 서로 연관지어 주며, 이러한 개체에 대해 표현한 단어들을 멘션(mention)이라 하며, 이런 멘션을 찾아내는 것을 멘션탐지(mention detection)라 한다. 멘션은 명사나 명사구를 기반으로 정의되며, 명사구의 경우에는 수식어를 포함하기 때문에 멘션탐지를 순차 데이터 문제(sequence labeling problem)로 정의할 수 있다. 순차 데이터 문제에는 Recurrent Neural Network(RNN) 종류의 모델을 적용할 수 있으며, 모델들은 Long Short-Term Memory(LSTM) RNN, LSTM Recurrent CRF(LSTM-CRF), Bidirectional LSTM-CRF(Bi-LSTM-CRF) 등이 있다. LSTM-RNN은 기존 RNN의 그레디언트 소멸 문제(vanishing gradient problem)를 해결하였으며, LSTM-CRF는 출력 결과에 의존성을 부여하여 순차 데이터 문제에 더욱 최적화 하였다. Bi-LSTM-CRF는 과거입력자질과 미래입력자질을 함께 학습하는 방법으로 최근에 가장 좋은 성능을 보이고 있다. 이에 따라, 본 논문에서는 멘션탐지에 Bi-LSTM-CRF를 적용할 것을 제안하며, 각 딥 러닝 모델들에 대한 비교실험을 보인다.

주제어: 멘션탐지, Bidirectional LSTM-CRF, 상호참조해결

1. 서론

최근 자연어처리에서는 딥 러닝을 이용한 의미분석 및 정보추출에 대한 연구가 진행 중이다(즉, 개체명 인식, 상호참조해결 등)[1,2]. 이 중에서 상호참조해결은 특정 개체에 대하여 다르게 표현한 단어들을 서로 연관지어주는 방법이며, 이런 개체에 대하여 표현한 단어 후보들을 멘션(mention)이라 정의한다. 문장 내에서 멘션을 찾아 정의하는 것을 멘션탐지(mention detection)라 한다. 각각의 멘션은 명사구를 기반으로 정의되고, 명사구에서 핵심이 되는 중심어(head)와 중심어를 수식하는 수식어로 이루어지므로, 멘션에 포함된 수식어 정보를 정확하게 찾아내야 멘션이 가리키는 개체가 명확해지고, 올바른 상호참조해결을 수행할 수 있게 된다. 멘션탐지는 순차 데이터 문제(sequence labeling problem)로 정의할 수 있으며, 본 논문에서는 멘션탐지를 위하여 딥 러닝(deep learning)의 방법 중 순차 데이터 문제에 적합한 Recurrent Neural Network(RNN) 종류의 모델들을 적용한다.

딥 러닝은 자연어처리, 음성인식, 패턴인식 등 여러 분야에서 각광받고 있다. 딥 러닝은 여러 층의 비선형 변환(non-linear activation)의 조합으로 입력 자질에 대하여 높은 수준의 추상화를 수행할 수 있으며, 이런 이유로 기존 기계학습 방법과 달리 자질 디자인과 자질 조합을 자동으로 수행할 수 있는 장점을 가진다[1]. 딥 러닝 모델 중 하나인 RNN은 순차 데이터를 모델링 할 수 있는 강력한 모델이다. 그러나 RNN은 그레디언트 소멸 문제(vanishing gradient problem)가 발생하는 문제가 있다[3, 4]. 이에 따라, Long Short-Term Memory(LSTM) 기반의 RNN은 히든 레이어(hidden layer)에 게이트(gate)를 정의하고, 가중치(weight) 행렬과 게이트들을 학습하여 필요한 부분에 에러율을 전파하는 방법으로 그레디언트 소멸 문제를 해결하였다[3]. 최근에는 LSTM-RNN의 출력 결과(output label)에 의존성(전이 확률)을 부여한 LSTM Recurrent CRF(LSTM-CRF)의 등장으로 순차 데이터 문제해결 성능이 더욱 향상되었으며[5], 이러한 LSTM-CRF를 전방향(forward) 모델 뿐만 아니라 후방향(backward) 모델을 함께 학습하여 POS 태깅, 개체명 등의 순차 데이터 문제에 대하여 우수한 성능을 보이고 있다[6].

본 논문에서는 Bidirectional LSTM CRF(Bi-LSTM-CRF)를 멘션탐지에 적용할 것을 제안하고, 각 RNN 모델들(즉, RNN, LSTM-RNN, LSTM-CRF, Bi-LSTM-CRF 등)에 대한 비교실험을 통하여 Bi-LSTM-CRF가 멘션탐지 문제에 적합하다는 것을 보인다.

2. 관련 연구

상호참조해결에서는 명사나 명사구로 정의된 멘션을 기반으로 참조해결을 수행한다. 이런 멘션들이 많을수록 상호참조해결의 가능성이 높아져 재현율이 향상되는 효과가 있다. 이처럼 임의의 문장에 대하여 멘션을 찾아내는 것을 멘션탐지라고 하며, 기존에는 규칙기반[7]과 통계기반[8]의 방법을 이용하여 적용되어 왔다.

규칙기반 멘션탐지는 의존트리(dependency tree)에 기반 하여 모든 명사와 명사구들을 추출하고 멘션으로 정의하게 된다. 이와 같은 방법으로 멘션을 추출하게 되면, 일반 단일 명사나 짧은 수식어를 포함한 명사(예를 들어, 형용사+중심어: 아름다운 꽃, 복합 명사: 한국전자통신연구소 등)와 같이 짧은 명사구는 정확하게 추출할 수 있으나, 규칙기반에서 수식어가 길어지는 경우에는 정확한 멘션으로 추출하기 어려우며, 그 외로 멘션

간의 크로스 문제, 오류 누적에 의한 문제 등이 발생할 수 있다[7].

앞서 언급한 멘션탐지에 대한 문제를 해결하기 위하여 [8]에서는 딥 러닝을 이용하여 멘션탐지를 수행하였다. 그러나 순차 데이터 문제인 멘션탐지에 Feed-Forward Neural Network(FFNN)를 적용하였기 때문에 높은 성능을 보이지 못했다. 이에 따라, 본 논문에서는 순차 데이터 문제에 적합한 RNN 모델들을 멘션탐지에 적용하여 비교 실험을 수행하고 Bi-LSTM-CRF가 멘션탐지에 가장 적합하다는 것을 보인다.

3. 멘션탐지

멘션탐지는 임의의 문장에서 멘션들을 추출해내는 것을 말하며, 멘션은 상호참조해결에서 어떤 개체를 가리킬 때 사용되는 후보 단어들로 정의된다. 모든 멘션은 명사 및 명사구를 기반으로 정의되고, 각각의 멘션은 하나 이상의 단어를 포함하게 되며, 본 논문에서는 이와 같은 순차 데이터 문제에 적합한 BIO 태그를 이용하여 딥 러닝에 적용하였다. [표 1]은 멘션탐지에 대한 예를 보인다.

표 1. 멘션탐지 예

입력문장
김대중 대통령의 프랑스 방문 중 한국 삼성자동차 인수를 공식 제의할지 모른다고 르노사의 한 관계자가 1일 밝혔다.
멘션탐지 결과
[[[김대중] 대통령의] [프랑스] 방문**]** 중 **[**[[한국] 삼성자동차] 인수를**]** **[**공식**]** 제의할지 모른다고 **[**[[르노사의] 한 관계자가**]** **[**1일**]** 밝혔다.

[표 1]은 모든 멘션이 명사와 명사구를 기반으로 추출됨을 보여주며, 모든 멘션들은 어절단위라고 가정한다. 여기서 멘션의 시작과 끝을 나타내는 인덱스를 바운더리라고 정의한다. [표 1]에서의 바운더리는 대괄호([])로 표현하며, 각각의 멘션 또는 안은 멘션(굵은 글씨 대괄호)과 안긴 멘션(일반 글씨 대괄호) 등을 구분할 수 있다. 안은 멘션은 하나의 명사구에서 멘션이 중복되어 정의되는 경우에 가장 넓은 바운더리를 가진 멘션을 의미하며, 안긴 멘션은 해당 멘션 안에 중복되어 정의된 멘션들을 의미한다. 예를 들어 "김대중 대통령의 프랑스 방문"은 [김대중 대통령의 프랑스 방문], [김대중 대통령의], [김대중], [프랑스]와 같은 멘션들로 정의되고, [김대중 대통령의 프랑스 방문]이 안은 멘션이며, 나머지 멘션들이 안긴 멘션이다.

위와 같이, 안은 멘션과 안긴 멘션에 대하여 BIO 태그로 표현하게 되면 서로 중복되는 문제가 발생한다. 따라서 본 논문에서는 가장 넓은 바운더리를 가지는 안은 멘션을 BIO 태그 표현의 기준으로 사용하였다.

4. Deep learning 모델

하나의 신경망은 입력 단어 x와 출력 레이블 y, 히든 레이어의 노드 h로 구성된다. 이를 기반으로 히든 레이어가 여러 층이 되는 구조를 딥 러닝이라 한다. 일반적인 FFNN은 고정된 입력과 출력에 대한 학습만 가능하기 때문에 순차 데이터에 적합하지 않다. 이에 따라, 본 논문에서는 순차 데이터에 적합한 RNN 모델들을 멘션탐지에 적용하였다.

4.1 Recurrent Neural Network

RNN은 [그림 1]과 같이 FFNN 구조를 입력 단어 열에 따라 확장한 모델로서, 이전 히든 레이어의 상태를 누적하여 사용하기 때문에 순차 데이터 문제에 적합하다. [그림 1]에서는 멘션탐지에 대하여 순차 데이터 문제로 표현한 것으로, BIO 태그를 이용한다. 여기서 B는 멘션의 시작 단어에 대한 표현이고, I는 멘션 내에 포함된 단어를 나타낸다. 그리고 O는 멘션에 포함되지 않는 단어를 뜻한다. 예를 들어, "김대중 대통령은 프랑스에 방문할 것이다(즉, 김대중/nnp 대통령/nng 은/jx 프랑스/nnp 에/jkb 방문/nng 하/xsv 르/etm 것/nnb 이/vcp 다/ef ./sf)"와 같은 문장은 각 형태소에 따라 [B I I B I O O O O O O O]와 같은 태그 열로 표현된다.

RNN의 입력 레이어(input layer)는 입력 단어 열 $x = (x_1, \ldots, x_T)$가 되며, 출력 레이어(output layer)는 출력 태그 열(멘션탐지 태그 열) $y = (y_1, \ldots, y_T)$가 된다. 이에 따라, RNN은 각 입력에 대한 신경망이 구성되며, 순차적 분류가 가능하게 되고, 다음과 같이 정의된다.

$$h_t = f(Ux_t + Vh_{t-1})$$
$$y_t = g(Wh_t)$$

위 식에서 U, V, W는 weight 행렬이며, f는 활성함수(sigmoid, tanh, relu 등), g는 softmax 함수이다.

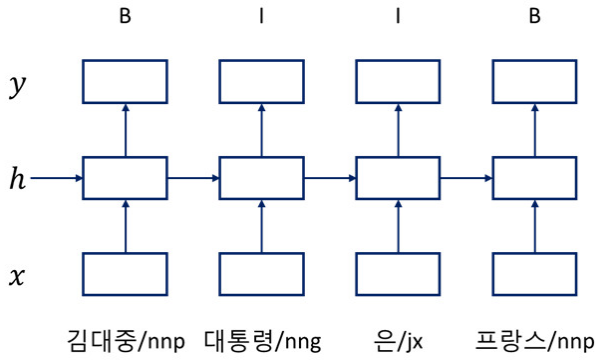

그림 1. RNN 모델

4.2. Long Short-Term Memory RNN

본 논문에서는 RNN의 그레디언트 소멸 문제를 해결하기 위하여 LSTM을 적용하였다. [그림 2]는 LSTM의 메모리 셀(memory cell)에 대한 구조를 나타낸다.

[그림 2]와 같이 각각의 메모리 셀들을 갱신하여 새로 입력되는 벡터와 기존의 셀(cell) 벡터의 값을 조정한다. 이에 따라, 멀리 떨어진 단어의 자질 값을 손실 없이 전달할 수 있어 RNN의 그레디언트 소멸 문제를 해결할 수 있다. LSTM-RNN은 다음과 같이 정의된다.

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_{t-1} + b_o)$$
$$h_t = o_t \odot \tanh(c_t)$$
$$y_t = g(W_{yh}h_t + b_y)$$

위 식에서 σ는 sigmoid 함수이며, $\odot$는 벡터간의 element-wise product이다. i, f, o, c는 각각 input gate, forget gate, output gate, memory cell이며, 모두 같은 히든 레이어 유닛 수를 가진다. W는 weight 행렬(예를 들어, W_{ih}는 input-hidden gate 행렬이고, W_{ox}는 output-input gate 행렬이다)이며, b는 bias term이다.

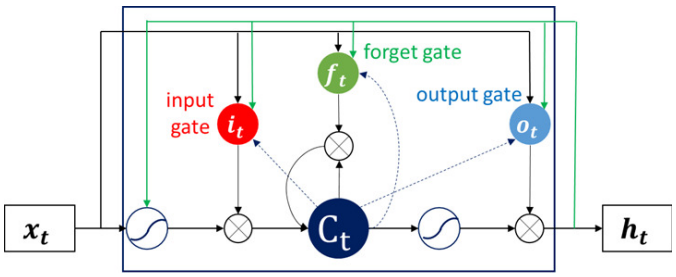

그림 2. Long Short-Term Memory Cell

4.3 LSTM Recurrent CRF

[그림 3]은 LSTM-CRF에 대한 구조를 나타낸다[5]. LSTM-RNN의 경우에는 각각 독립적인 출력 결과를 갖는 모델인 반면, LSTM-CRF는 각 출력 결과에 의존성(전이 확률)이 부여된 모델이다. LSTM-CRF는 LSTM의 특징인 이전 시간의 상태를 누적하는 것과 CRF의 특징인 문장 단위의 태그 정보를 적용할 수 있는 것 때문에 순차 데이터 문제에 보다 더 좋은 성능을 보일 수 있다. LSTM-CRF의 output layer의 식은 다음과 같이 확장된다.

$$y = W_{yh}h_t + b_y$$
$$s(x,y) = \sum_{t=1}^{T} A(y_{t-1}, y_t) + y_t$$
$$\log p(y|x) = s(x,y) - \log \sum_{y'} \exp(s(x,y'))$$

위 식에서 $A(y_{t-1}, y_t)$는 이전 출력 태그 y_{t-1}에서 다음 출력 태그 y_t로 전이될 확률을 의미하는 함수이고, $s(x,y)$는 멘션탐지 태그열 y의 점수를 구하는 score 함수이다. 여기서 $\log p(y|x)$를 구하기 위하여 CRF와 같이 forward 알고리즘을 이용하며, 최적의 태그열을 찾기 위해 viterbi search 알고리즘을 이용한다.

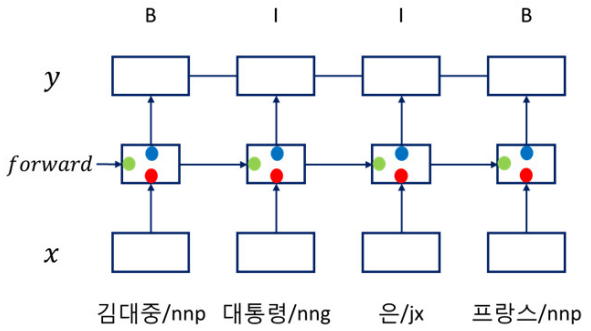

그림 3. LSTM Recurrent CRF 모델

4.4 Bidirectional LSTM CRF

본 논문에서 제안한 Bidirectional LSTM CRF(Bi-LSTM-CRF)는 앞서 말한 LSTM-CRF를 전방향(forward) 학습뿐만 아니라 후방향(backward) 학습도 함께 수행하는 모델이다. Bi-LSTM-CRF는 과거입력자질(past input feature)에 대한 추상화뿐만 아니라 미래입력자질(future input feature)에 대한 추상화를 모두 적용할 수 있기 때문에 Bi-LSTM-CRF는 양방향 모두의 자질 조합 및 추상화 정보를 얻을 수 있다.

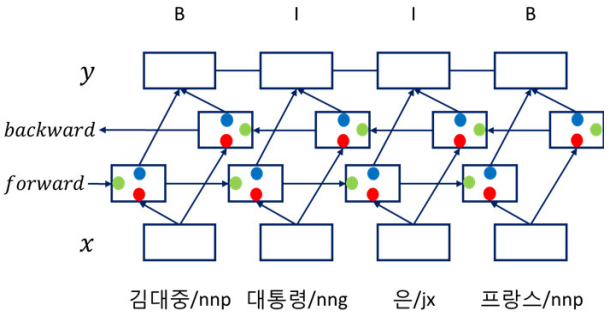

그림 4. Bidirectional LSTM-CRF

본 논문에서는 Bi-LSTM-CRF 모델의 학습을 수행하기 위하여 Stochastic Gradient Descent(SGD)와 Back-Propagation Through Time(BPTT) 알고리즘을 이용한다.

5. 실험

실험에 사용된 학습 데이터는 뉴스 도메인 150 문서, 퀴즈 도메인 350 쌍으로 총 500 문서이다. 이 중, 학습 데이터는 뉴스 도메인 140 문서, 퀴즈 도메인 310 쌍을 사용하였고, 테스트 데이터로는 뉴스 도메인 10 문서, 퀴즈 도메인 40 쌍을 사용하였다.

본 논문에서 제안한 Bi-LSTM-CRF 모델을 수행하기 위

하여 적용한 자질은 [8]에서 사용한 자질과 같다. 자질 표현(feature embedding)에는 평균 0과 분산 0.01이 되도록 무작위로 초기화 시킨 값을 이용하였으며, 단어 표현(word embedding)은 10만 단어에 대한 세종 테그셋을 Neural Network Language Model(NNLM)로 학습한 것을 사용하였다. 그리고 성능 측정을 위한 척도(measure)는 F1 값을 사용하였으며, 본 논문에서 제안한 딥 러닝 모델들(즉, FFNN, RNN, LSTM, LSTM-CRF, Bi-LSTM-CRF 등)을 이용하여 비교실험을 수행하였다. 또한 각 모델들은 1-layer LSTM으로 학습을 수행하였으며, 학습율(learning rate)은 0.1을 시작으로 성능 개선이 없으면 3 에포크(epoch)마다 50%씩 감소하도록 정의하여 학습하였다. 투사 레이어(projection layer)와 히든 레이어(hidden layer)에는 drop-out을 각각 0.2와 0.5의 확률 값으로 적용하였고, 각 레이어에 대한 활성함수는 sigmoid와 tanh를 모두 고려하여 실험을 수행하였다(FFNN은 relu도 적용하였다).

[표 2]는 본 논문에서 제안한 각 딥 러닝 모델에 따른 멘션탐지 실험 결과이다.

표 2. 멘션탐지 실험 결과 (F1)

모델	F1
FFNN (relu)	68.57
RNN (tanh)	65.91
LSTM-RNN (tanh)	66.51
LSTM-CRF (tanh)	73.67
Bi-LSTM-CRF (sigm)	**76.24**

실험 결과, FFNN은 히든 레이어의 유닛 수가 300이고 relu 함수를 적용하였을 때 68.57%의 성능을 보였다. RNN과 LSTM-RNN은 tanh 함수를 적용할 경우에 각각 65.91%, 66.51%를 보였다. 여기서 LSTM-RNN이 RNN보다 높은 성능을 보였으나 두 모델 모두 FFNN보다 낮은 성능을 보였다. LSTM-CRF는 히든 레이어 유닛 수가 300이고, tanh 함수를 사용하였을 때 73.67%로 이전 모델들에 비하여 좋은 성능을 보였다. 이것은 출력 결과 층(layer)에 대하여 의존성을 부여함으로 문장 단위 정보를 포함하기 때문인 것으로 보인다. 그리고 마지막으로 Bi-LSTM-CRF는 히든 레이어 유닛 수가 500이고, sigmoid 함수를 이용하였을 때, FFNN에 비하여 약 7.67% 향상되었고, LSTM-CRF보다 약 2.57% 향상되어 76.24%의 성능으로 모델들 중에서 가장 높은 성능을 보였다. Bi-LSTM-CRF는 각 단어의 과거(forward)와 미래(backward) 입력 자질에 대한 추상화를 이용하여 학습을 수행하는데, 이 방법이 순차 데이터 문제로 정의한 멘션탐지에 적합하다는 것을 알 수 있다.

6. 결론

본 논문에서는 상호참조해결의 참조해결을 위하여 기반이 되는 멘션탐지에 대하여 순차 데이터 문제에 적합한 RNN 종류의 모델들(즉, RNN, LSTM-RNN, LSTM-CRF, Bi-LSTM-CRF 등)을 적용할 것을 제안하였다. 각 모델별로 비교 실험을 수행한 결과, FFNN이 약 68.57%, LSTM-CRF가 약 73.67%, 그리고 Bi-LSTM-CRF가 약 76.24%의 성능을 보였다. 이 중에서 Bi-LSTM-CRF는 FFNN에 비하여 약 7.67%의 향상된 성능을 보였으며, Bi-LSTM-CRF가 LSTM-CRF에 비하여 멘션탐지에 적합하다는 것을 알 수 있었다.

향후 연구로는 본 논문에 따른 Bi-LSTM-CRF를 이용한 멘션 탐지 방법을 상호참조해결에 적용할 것이며, 상호참조해결에도 LSTM 모델을 적용할 예정이다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임. (No.R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] R. Collobert, et al. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12, 2011.

[2] 박천음, 최경호, 이창기. 딥 러닝을 이용한 가이드 멘션페어 한국터 상호참조해결, *한국정보과학회 2015 한국컴퓨터종합학술대회 논문집*, pp. 693-695, 2015

[3] Kaisheng Yao, et al. Spoken language understanding using long short-term memory neural networks. *In Spoken Language Technology Workshop (SLT), 2014 IEEE* pp.189-194. IEEE, 2014.

[4] Kaisheng Yao, et al. Recurrent conditional random field for language understanding, In Acoustics, *Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* pp. 4077-4081. IEEE, 2014.

[5] 이창기, Long short-term memory 기반의 recurrent neural network를 이용한 개체명 인식, *한국정보과학회 2015 한국컴퓨터종합학술대회 논문집*, pp. 645-647, 2015.

[6] Zhiheng Huang, et al. Bidirectional lstm-crf models for sequence tagging, *arXiv preprint arXiv: 1508.01991,* 2015.

[7] 박천음, 최경호, 이창기. Multi-pass Sieve를 이용한 한국어 상호참조해결. *정보과학회논문지 41.11*, pp. 992-1005, 2014.

[8] 박천음, 이창기, 딥 러닝을 이용한 상호참조해결 멘션 탐지, 제9회 한국정보과학회 한국빅데이터학회 공동학술 심포지엄, 2015.