

Backward LSTM CRF를 이용한 한국어 의미역 결정

배장성[○], 이창기, 임수중⁺

강원대학교 컴퓨터학과, 한국전자통신연구원⁺

jseffort88@gmail.com, leeck@kangwon.ac.kr, lsj@etri.re.kr

Korean Semantic Role Labeling using Backward LSTM CRF

Jangseong Bae[○], Changki Lee, Soojong Lim⁺
Kangwon National University, ETRI⁺

요 약

Long Short-term Memory Network(LSTM) 기반 Recurrent Neural Network(RNN)는 순차 데이터를 모델링 할 수 있는 딥 러닝 모델이다. 기존 RNN의 그라디언트 소멸 문제(vanishing gradient problem)를 해결한 LSTM RNN은 멀리 떨어져 있는 이전의 입력 정보를 볼 수 있다는 장점이 있어 음성 인식 및 필기체 인식 등의 분야에서 좋은 성능을 보이고 있다. 또한 LSTM RNN 모델에 의존성(전이 확률)을 추가한 LSTM CRF모델이 자연어처리의 한 분야인 개체명 인식에서 우수한 성능을 보이고 있다. 본 논문에서는 한국어 문장의 지배소가 문장 후위에 나타나는 점에 착안하여 Backward 방식의 LSTM CRF 모델을 제안하고 이를 한국어 의미역 결정에 적용하여 기존 연구보다 더 높은 성능을 얻을 수 있음을 보인다.

주제어: 한국어 의미역 결정, Backward, LSTM CRF

1. 서론

자연어처리 모듈 개발에 사용되는 대부분의 기계학습 알고리즘들은 사람이 고안한 자질(feature)을 입력으로 받고 이 자질들의 최적의 가중치(weight)를 구한다. 그러나 각 자연어처리 모듈마다 적합한 자질을 설계하고 최적의 자질 조합을 구하는 것은 많은 시간과 노력을 필요로 한다. 이러한 문제를 해결하기 위해 자질들을 높은 수준의 표현으로 추상화 시켜줄 수 있는 딥 러닝(deep learning) 기술이 최근 많이 연구되고 있다[1]. 딥 러닝은 비선형(non-linear)의 Hidden layer가 여러 층으로 쌓여 이루어진 인공신경망(artificial neural network)으로, 입력 자질들을 여러 비선형 변환기법의 조합을 통해 높은 수준의 표현으로 추상화할 수 있는 장점이 있다. Long Short-term Memory(LSTM)를 이용한 Recurrent Neural Network(RNN)는 기존 RNN 모델의 그라디언트 소멸(vanishing gradient problem) 문제[2]를 해결한 딥 러닝 모델이다. LSTM 기반 RNN은 음성 인식, 필기체 인식, 언어모델, 자연어 이해 등의 분야에서 우수한 성능을 보이고 있다[2].

의미역(semantic role)은 문장 내에서 서술어에 의해 기술되는 행동이나 상태에 대한 명사구의 의미 역할을 말하며 의미역이 부여된 각 명사구를 논항(argument)이라고 한다. 의미역 결정(semantic role labeling)은 문장의 각 서술어의 의미와 그 논항들의 의미역을 결정하여 “누가, 무엇을, 어떻게, 왜” 등의 의미 관계를 찾아내는 자연어처리의 한 단계이다. 최근 의미역 결정 연구에는 기계학습 알고리즘 중 하나인 Structural SVM이 좋은 성능을 보이고 있다[3]. [4]에서는 딥 러닝 기술을 한국어 의미역 결정에 적용하여 자질 설계 및 자질 조합의 최적화 없이 Structural SVM 기반 모델과 비슷한 성능을 얻고자 했으나, 이전 단어의 정보를 볼 수 없는 Feed Forward Neural Network(FFNN) 모델을 사용하여 그

성능이 기존 Structural SVM에 미치지 못하였다.

본 논문에서는 한국어 의미역 결정을 sequence labeling 문제로 보고 순차 데이터(sequential data) 모델링에 적합한 LSTM RNN을 한국어 의미역 결정에 적용하고, [5]에서 제안한 LSTM Conditional Random Field(CRF)를 사용하여 의미역 태그 사이의 의존성(전이 확률)을 추가한다. 또한 한국어 문장의 지배소라 할 수 있는 서술어가 문장의 뒤쪽에 나타나는 점에 착안하여 Backward LSTM CRF를 제안하고, 한국어 의미역 결정에 적용하여 기존 Structural SVM 및 Forward 방식의 LSTM CRF 모델보다 높은 성능을 얻을 수 있음을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 Backward LSTM CRF 모델에 대해 설명하고, 4장에서는 실험 및 결과를 분석한다. 5장에서는 결론에 대해 기술한다.

2. 관련 연구

의미역 결정 연구는 크게 격렬사전에 기반을 둔 방법과 말뭉치에 기반을 둔 방법으로 나눌 수 있다. 격렬사전에 기반을 둔 방법은 서술어와 논항들의 쓰임을 기술한 격렬사전을 이용하는 방법으로, 서술어와 논항에 대한 문법 관계를 기술한 격렬(frame)과 논항들의 정보를 기술한 선택제약(selectional restriction) 등을 이용하여 서술어-논항 관계에 부합하는 격렬을 선택하여 의미역을 결정하는 방법이다. 격렬사전에 기반을 둔 방법은 입력 문장과 격렬 사이의 유사도 계산 과정을 통해 의미역이 결정되기 때문에 처리속도가 빠르고 높은 정확률을 보이지만, 격렬사전의 구축이 어렵고 격렬사전에 기술되지 않은 임의격을 처리하지 못하는 문제가 있다[6].

말뭉치에 기반을 둔 방법은 의미역이 태깅된 말뭉치를 구축하고 이를 이용하여 기계학습 방법으로 의미역을 결정하는 방법이다. 이 방법은 격렬사전에 기반을 둔 방법

에 비해 적용률이 높은 장점이 있으나, 의미역이 태깅된 말뭉치의 구축이 어렵다는 단점이 있다[6]. 최근에는 의미역 말뭉치와 Structural SVM과 같은 기계학습을 이용한 연구가 활발히 이루어지고 있다[3].

[4]에서 사용한 Feed Forward Neural Network 모델은 출력 레이블을 결정하기 위해 현재 입력 단어를 포함한 고정된 크기의 윈도우만 볼 수 있다는 단점이 있다. [5]에서는 sequential data 모델링에 적합한 LSTM RNN 모델에 출력 레이블간의 의존성을 추가하고 Viterbi search를 통한 최적의 출력 레이블 열을 찾아주는 LSTM CRF 모델을 제안하였고 기존 방법들보다 높은 성능을 얻었다. LSTM CRF 모델은 LSTM 구조로 인해 멀리 떨어져 있는 이전 단어의 정보를 볼 수 있는 장점이 있지만 뒤에 나올 단어의 정보를 볼 수 없는 단점 또한 존재한다. 대부분의 한국어 문장 지배소가 문장 후위에 나타나는 점을 고려하면 LSTM CRF를 한국어 의미역 결정에 적용 시 그 효과가 미미할 수 있다. 따라서 본 논문에서는 Backward LSTM CRF를 제안하고 한국어 의미역 결정에 적용한다.

3. Backward LSTM CRF 모델

RNN은 sequential data를 처리하는데 적합한 형태로 디자인 되어 있으며 RNN을 unfold한 구조는 그림 1과 같다. 입력 단어 열 $x = (x_1, x_2, \dots, x_t)$ 와 Hidden layer의 유닛 열 $h = (h_1, h_2, \dots, h_t)$, 출력 단어 열(의미역 태그 열)을 $y = (y_1, y_2, \dots, y_t)$ 라 할 때 RNN은 아래와 같이 정의된다.

$$h(t) = f(Ux(t) + Wh(t-1))$$

$$y(t) = g(Vh(t))$$

위 식에서 U, W, V 는 가중치 행렬이며 $f(z)$ 는 Sigmoid 혹은 Tanh 함수이고 $g(z)$ 는 Softmax 함수이다.

$$f(z) = \frac{1}{1+e^{-z}} \text{ or } \frac{1-e^{-2z}}{1+e^{-2z}}$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

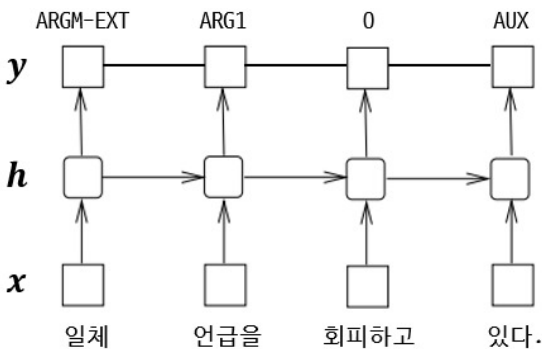


그림 1. RNN 구조

본 논문에서 한국어 의미역 결정에 적용하는 LSTM은 다음과 같이 정의된다.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o)$$

$$h_t = o_t \tanh(c_t)$$

위 식에서 σ 는 sigmoid 함수이고, i, f, o, c 는 각각 input gate, forget gate, output gate, memory cell vector (벡터)이며 각 벡터의 크기는 Hidden layer 벡터 크기와 같다. weight(가중치) 행렬의 아래첨자는 연결된 각 노드를 표시해 준다. 예를 들어 W_{hi} 는 Hidden layer와 input gate간의 가중치 행렬이다. 그림 2는 LSTM memory cell의 구조를 나타낸다. LSTM의 Hidden layer는 그림 2의 memory cell에 의해 업데이트되기 때문에 이전 단어들의 정보와 현재 단어의 정보를 손실 없이 유지할 수 있어 그라디언트 소멸 문제에서 자유롭다.

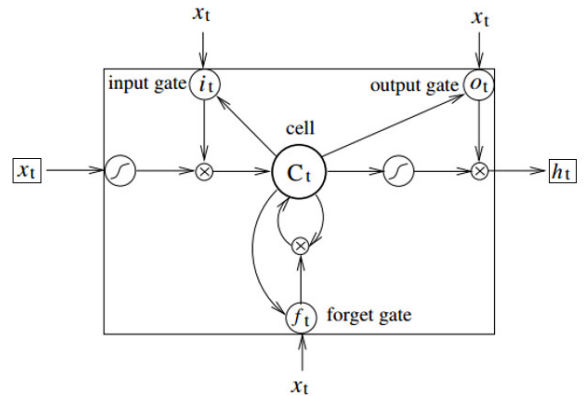


그림 2. LSTM memory cell 구조

본 논문에서는 Conditional Random Field(CRF)를 LSTM에 추가한 [5]의 모델을 한국어 의미역 결정에 적용한다. CRF는 레이블의 인접성에 대한 정보를 바탕으로 현재 레이블을 추측하는 방법이다. [5]에서는 출력 레이블간의 의존성을 추가하기 위해 output layer를 다음과 같이 확장하였다.

$$y(t) = W_{hy}h_t + b_y$$

$$s(x, y) = \sum_{t=1}^T A(y_{t-1}, y_t) + y_t$$

$$\log p(y|x) = s(x, y) - \log \sum_y \exp(s(x, y'))$$

위 식에서 $A(y_{t-1}, y_t)$ 는 의미역 태그 y_{t-1} 에서 y_t 으로 전이될 확률을 의미하고, $s(x, y)$ 는 의미역 태그 열의 점수이다. $\log p(y|x)$ 를 구하기 위해서 CRF와 유사하게

forward 알고리즘을 이용하며, 최적의 태그 열을 구하기 위해 Viterbi search 알고리즘을 적용한다. 그림 3은 Backward LSTM CRF 모델의 구조를 나타낸다. LSTM RNN에 전이확률이 추가되고 입력 시퀀스가 역방향임을 알 수 있다.

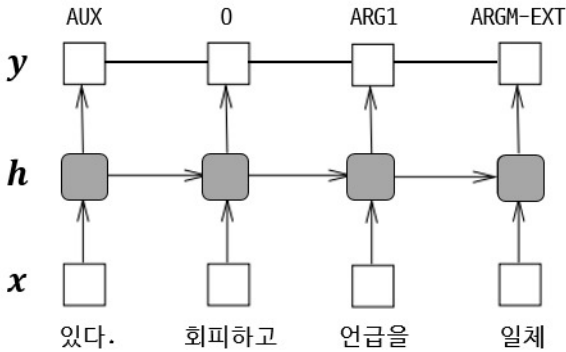


그림 3. Backward LSTM CRF 구조(회색박스: LSTM)

Backward LSTM CRF 모델의 학습을 위해 Stochastic Gradient Descent(SGD)를 이용하여 $-\log p(y|x)$ 를 최소화 시켰고, Back-Propagation Through Time(BPTT) 알고리즘을 이용하였다. 또한 과적합 문제를 줄이기 위해 Dropout 기술[7]을 적용하였다.

4. 실험

본 논문에서는 Backward LSTM CRF 모델을 기존 연구와 비교하기 위하여 기존 연구에서 사용한 Korean PropBank[8]를 학습 말뭉치로 사용하였으며, 기존 연구와 동일한 학습데이터, 평가 데이터를 구성하였다. 실험에 사용한 한국어 word embedding(단어 표현)은 word2vector[9]를 이용하여 구한 것을 사용하였다. feature embedding은 랜덤으로 초기화한 값을 사용하였고(평균 0, 분산 0.01), 또한 Projection layer와 Hidden layer에 Dropout 기술을 적용하였다(0.2, 0.5).

실험 결과의 성능 지표는 정확도와 재현율의 조화평균인 F1 지표를 사용하였으며, 본 논문에서 제시하고 있는 성능은 의미역 결정 문제의 논항 인식 및 분류(AIC)에 해당한다.

표 1은 Korean PropBank를 이용한 한국어 의미역 결정 실험 결과이다. 기본 자질과 word cluster 자질을 이용한 Structural SVM은 76.96%의 성능을 보였고, FFNN은 76.20%의 낮은 성능을 보였는데 이는 이전 단어의 정보를 볼 수 없기 때문이라고 볼 수 있다.

표 1. 한국어 의미역 결정 실험 결과(AIC)

모델	F1
Structural SVM	76.96
FFNN	76.20
Forward LSTM RNN	75.50
Backward LSTM RNN	76.81
Forward LSTM CRF	76.12
Backward LSTM CRF	77.21

Backward LSTM RNN 모델은 76.81%의 성능을 보여 FFNN보다 0.61%의 성능 향상이 있었고 본 논문에서 제안한 Backward 방식과 [5]에서 제안한 LSTM CRF를 적용한 Backward LSTM CRF 모델은 77.21%의 가장 높은 성능을 보였다. 반면 Forward 방식의 LSTM RNN과 LSTM CRF는 Backward 방식의 모델보다 각각 1%이상 성능이 하락하였다. 이를 통해 문장 지배소가 주로 문장의 후위에 나타나는 한국어에서 Backward 방식이 Forward 방식보다 더 유리함을 알 수 있다. 또한 LSTM RNN과 LSTM CRF의 성능 비교를 통해 출력 레이블간의 의존성을 잘 모델링할 경우 한국어 의미역 결정 성능 향상에 도움이 됨을 알 수 있다.

5. 결론

본 논문에서는 LSTM RNN 모델에 출력 레이블간의 의존성을 추가시킨 LSTM CRF를 한국어 문장의 특성을 맞게 변경한 Backward LSTM CRF를 제안하고, 이를 한국어 의미역 결정에 적용하여 기존 Structural SVM, FFNN 등의 모델보다 더 높은 성능을 얻었으며, 한국어 의미역 결정에 Forward 방식보다 Backward 방식이 더 유리함을 알 수 있었다.

향후 연구로는 LSTM RNN의 구조를 역방향으로 한층 더 쌓아 이전 단어의 정보 뿐 만 아니라 다음 단어의 정보를 함께 이용할 수 있는 Bidirectional LSTM을 한국어 의미역 결정에 적용해볼 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] Ronan Collobert, et al. Natural Language Processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493-2537. 2011.
- [2] YAO, Kaisheng, et al. Spoken language understanding using long short-term memory neural networks. In: Spoken Language Technology Workshop (SLT), IEEE. 189-194. 2014.
- [3] 이창기, 임수중, 김현기. Structural SVM 기반의 한국어 의미역 결정. 정보과학회논문지 제42권 제2호, 220-226. 2015.
- [4] 배장성, 이창기, 임수중. 딥 러닝을 이용한 한국어 의미역 결정. 한국컴퓨터종합학술대회 논문집. 690-692. 2015.
- [5] 이창기. Long Short-term Memory 기반의 Recurrent Neural Network를 이용한 개체명 인식. 한국컴퓨터종합학술대회 논문집. 645-647. 2015.
- [6] 김병수, 외 4인. 부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정. 한국컴퓨터종합학술대회, vol. 33, no. 1, 2006.

- [7] G.E Dahl, et al. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Acoustics, Speech and Signal Processing (ICASSP), International Conference on IEEE. p. 8609-8613. 2013.
- [8] Martha Palmer, et al. Korean Propbank. <http://catalog.ldc.upenn.edu/LDC2006T03>.
- [9] Tomas Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.