Highway BiLSTM-CRFs 모델을 이용한 한국어 의미역 결정

배장성*O, 이창기*, 김현기+

강원대학교*, 한국전자통신연구원+

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

Korean Semantic Role Labeling with Highway BiLSTM-CRFs

Jangseong Bae*O, Changki Lee*, Hyunki Kim+

Kangwon National University*, ETRI+

요 약

Long Short-Term Memory Recurrent Neural Network(LSTM RNN)는 순차 데이터 모델링에 적합한 딥러닝 모델이다. Bidirectional LSTM RNN(BiLSTM RNN)은 RNN의 그래디언트 소멸 문제(vanishing gradient problem)를 해결한 LSTM RNN을 입력 데이터의 양 방향에 적용시킨 것으로 입력 열의 모든 정보를 볼 수 있는 장점이 있어 자연어처리를 비롯한 다양한 분야에서 많이 사용되고 있다. Highway Network는 비선형 변환을 거치지 않은 입력 정보를 히든레이어에서 직접 사용할 수 있게 LSTM 유닛에 게이트를 추가한 딥러닝 모델이다. 본 논문에서는 Highway Network를 한국어 의미역 결정에 적용하여 기존 연구 보다 더 높은 성능을 얻을 수 있음을 보인다.

주제어: 자연어처리, 의미역 결정, 딥러닝, Highway Network

1. 서론

의미역은 서술어에 의해 기술되는 행동이나 상태에 대한 명사구의 의미 역할을 말하며 의미역이 부여된 각 명사구를 논항 이라고 한다. 의미역 결정은 각 서술어의 의미와 그 논항들의 의미역을 결정하여 “누가, 무엇을, 어떻게, 왜” 등의 의미 관계를 찾아내는 자연어처리의 한 응용이며 정보 추출, 질의 응답과 같은 다양한 자연어처리 시스템의 성능 향상을 위한 입력 정보로 사용될 수 있다. 예를 들어 의미역으로부터 시간 및 공간 정보, 사건의 주체와 같이 문장이 가지는 의미 등을 파악해 질의 응답 시스템이 필요로 하는 정보를 제공할 수 있다. 최근 의미역 결정 연구에는 Recurrent Neural Network(RNN)와 같은 딥러닝 모델을 이용한 연구가 주로 이루어 지고 있다[1,2,3].

딥러닝은 비선형의 히든레이어가 여러 층으로 쌓인 인공 신경망으로, 입력 자질들을 여러 비선형 변환기법의 조합을 통해 높은 수준의 표현으로 나타낼 수 있는 장점이 있다. Long Short-Term Memory Recurrent Neural Network(LSTM RNN)는 기존 RNN 모델의 그래디언트 소멸 문제(vanishing gradient problem)[4]를 해결한 딥러닝 모델이다. LSTM RNN은 음성 인식, 기계 번역, 자연어이해 등 다양한 분야에서 우수한 성능을 보이고 있으며[5,6], 순차 데이터(sequential data) 모델링에 적합한 구조로 이루어져 있다. Highway Network[7,8]는 비선형 변환을 거친 정보를 사용하는 LSTM 유닛에 비선형 변환을 거치지 않은 정보를 일부 선택하여 볼 수 있게 설계된 딥러닝 모델이다. 본 논문에서는 Highway Network를 한국어 의미역 결정에 적용하여 기존 연구보다 더 좋은 성능을 나타냄을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 소개하고, 3장에서는 한국어 의미역 결정 모델에 대해 설명하고, 4장에서는 실험 및 결과를 분석한다. 5장에서는 결론에 대해 기술한다.

2. 관련연구

최근 의미역 결정 연구는 딥러닝을 이용한 연구가 주로 이루어 지고 있다. [1]에서 사용한 Feed Forward Neural Network 모델은 출력 레이블을 결정하기 위해 현재 입력 단어를 포함한 고정된 크기의 입력 정보를 사용하는 단점이 있다. [2]의 연구는 Bidirectional Long Short-Term Memory Recurrent Neural Network(BiLSTM RNN) 모델을 이용한 연구로 LSTM 구조를 통해 멀리 떨어져 있는 단어의 정보를 유지할 수 있는 장점이 있고, Bidirectional 방법을 이용하여 문장에 나타나는 모든 단어의 정보를 사용하였다. [3]의 연구는 이미지 인식에서 뛰어난 성능을 보이고 있는 Convolutional Neural Network(CNN) 모델을 이용하여 의미역 결정에 사용되는 입력을 재구성하고, 재구성된 입력을 LSTM RNN의 입력으로 사용한다. [3]은 기존 단어 단위 연구가 아닌 알파벳 단위의 연구로서 입력 단어의 Out of vocabulary 문제에서 자유롭고 새로운 단어 표현을 얻는 장점이 있다. [7,8]은 비선형 변환을 거친 입력 정보를 사용하는 LSTM RNN 모델에 비선형 변환을 거치지 않은 정보를 사용할 수 있게끔 LSTM 유닛에 게이트를 추가한 딥러닝 모델이다. 본 논문에서는 [7,8]의 모델을 한국어 의미역 결정에 적용한다.

3. Highway BiLSTM-CRFs 모델

RNN은 순차 데이터 모델링에 적합한 형태로 디자인 되어 있으며 RNN을 입력 순서에 따라 언폴드(unfold)한 구조는 그림 1과 같다. 입력 단어 열 $x = \{x_1, x_2, \dots, x_T\}$와

히든레이어 유닛 열 $h=\{h_1,h_2,\dots,h_T\}$, 출력 단어 열을 $y=\{y_1,y_2,\dots,y_T\}$라 할 때 RNN은 식 (1)과 같이 정의된다.

$$h_t = f(UEx_t + Vh_{t-1} + b_h) \quad (1)$$
$$P(y_t|x) = y_t{}^T g(Wh_t + b_y)$$

U,W,V는 가중치 행렬이며 E는 단어 및 자질의 가중치 행렬이다. $f(z)$는 Sigmoid 혹은 Tanh 함수이고 $g(z)$는 Softmax 함수이다.

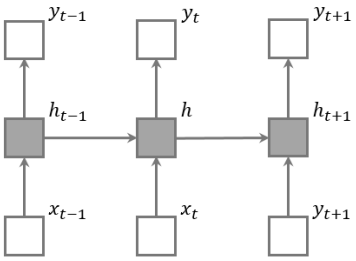

그림 1. RNN 모델

RNN은 입력 데이터의 열이 길어지면 신경망 구조가 깊어져 에러 전파(error propagation)가 어려워지는 그래디언트 소멸 문제가 발생하는데, 이 문제를 해결한 LSTM RNN은 식 (2)와 같이 정의된다.

$$i_t = \sigma(W_{ix}Ex_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (2)$$
$$f_t = \sigma(W_{fx}Ex_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}Ex_t + W_{ch}h_{t-1} + b_c)$$
$$o_t = \sigma(W_{ox}Ex_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o)$$
$$h_t = o_t \odot \tanh(c_t)$$
$$P(y_t|\boldsymbol{x}) = y_t{}^T g(W_{yh}h_t + b_y)$$

위 식에서 σ는 Sigmoid 함수이고, $\odot$는 벡터 간의 element-wise product를 나타낸다. i,f,o,c는 각각 input gate, forget gate, output gate, memory cell 벡터 이며 각 벡터의 크기는 히든레이어의 벡터 크기와 같다. 가중치 행렬의 아래 첨자는 연결된 각 노드를 표시해 준다. 예를 들어 W_{hi}는 히든레이어와 input gate간의 가중치 행렬이다. 그림 2는 LSTM 유닛의 구조를 나타낸다.

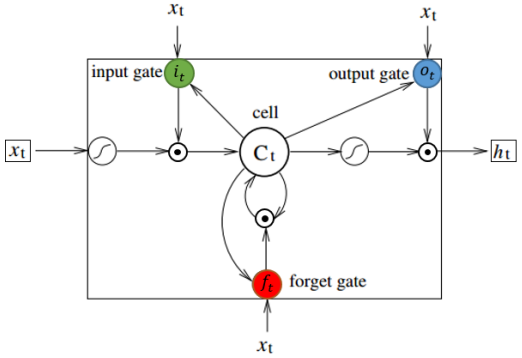

그림 2. LSTM 유닛 구조

그림 3은 LSTM RNN 구조를 나타내며 빨강, 파랑, 초록으로 표시된 부분이 LSTM RNN의 각 게이트(gate)를 나타낸다. LSTM의 입력은 4장의 표 1의 자질들이 projection layer를 거친 후 concatenate되어 하나의 벡터로 만들어지고, 만들어진 1개의 벡터가 LSTM의 입력으로 들어가게 된다.

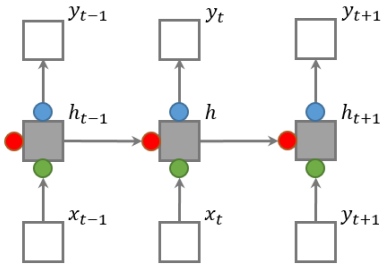

그림 3. LSTM RNN 모델

Highway Network는 비선형 변환을 거치지 않은 입력 정보를 LSTM 유닛이 사용할 수 있게 LSTM 유닛에 새로운 게이트를 추가한 모델이다. 추가된 게이트r_t와 변경된 h_t의 수식은 식 (3)과 같다.

$$r_t = \sigma(W_{rx}Ex_t + W_{rh}h_{t-1} + W_{rc}c_{t-1} + b_r) \quad (3)$$
$$h_t = r_t \odot o_t \odot \tanh(c_t) + (1 - r_t) \odot W_{hx}Ex_t$$

r_t는 비선형 변환을 거치지 않은 정보와 비선형 변환을 거친 정보를 얼마나 사용할지 정하는 역할을 하게 된다. 본 논문에서는 Highway Network를 BiLSTM RNN 모델에 적용하여 문장 전체의 정보를 사용한다. 또한 현재의 의미역 태그를 결정하기 위해 인접한 의미역 태그 정보를 활용하고자 한다. 이를 위해 출력 레이블의 인접성 정보를 바탕으로 현재 레이블을 추측할 수 있는 Conditional Random Field(CRFs)를 이용하여 output layer를 식 (4)와 같이 확장하였다.

$$s_{word}(y_t,t) = y_t{}^T(W_{yh}h_t + b_y) \quad (4)$$
$$s_{sent}(\boldsymbol{x},\boldsymbol{y}) = \sum_{t=1}^{T}\{[A]_{y_{t-1},y_t} + s_{word}(y_t,t)\}$$
$$log\,P(\boldsymbol{y}|\boldsymbol{x}) = s_{sent}(\boldsymbol{x},\boldsymbol{y}) - log\sum_{y'} exp(s_{sent}(\boldsymbol{x},\boldsymbol{y}'))$$

식 (4)에서 $[A]_{y_{t-1},y_t}$는 의미역 태그 y_{t-1}에서 y_t로 전이될 확률을 의미하고, $s_{sent}(\boldsymbol{x},\boldsymbol{y})$는 의미역 태그 열$y$의 점수이다. $log\,P(\boldsymbol{y}|\boldsymbol{x})$를 구하기 위해 forward 알고리즘을 이용하며, 최적의 태그 열을 구하기 위해 Viterbi search 알고리즘을 적용한다. 그림 4는 Highway BiLSTM-CRFs 모델을 나타낸다. **그림 4의 주황색 네모 상자는 비선형 변환을 거치지 않은 정보와 비선형 변환을 거친 정보를 얼마나 사용할지 정하는 r_t 게이트를 나타낸다.** 또한 모델의 출력 레이블 간의 의존성(전이확률)이 추가된 것을 알 수 있다. Highway BiLSTM-CRFs 모델의 학습을 위해 Stochastic Gradient Descent(SGD)를 이용하여 $-log\,p(y|x)$를 최소화 하였고, Back-Propagation Through Time(BPTT) 알고리즘을 이용하였다.

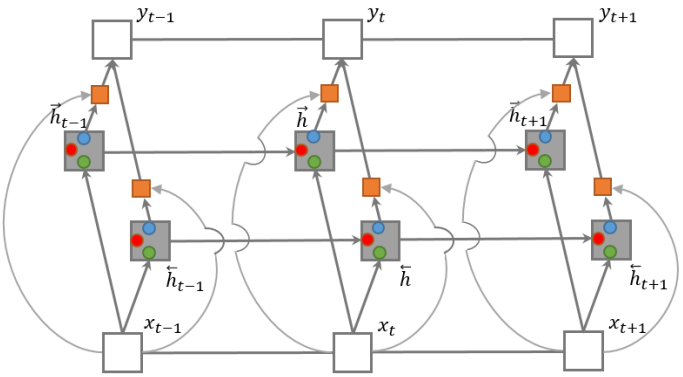

그림 4. Highway BiLSTM-CRFs 구조

4. 실험

본 논문에서는 Highway BiLSTM-CRFs 모델을 기존 연구와 비교하기 위하여 기존 연구에서 사용한 Korean PropBank[9]를 학습 말뭉치로 사용하였으며, 기존 연구와 동일한 학습데이터, 평가 데이터를 구성하였다. 표 1은 본 논문에서 사용하는 한국어 의미역 결정 자질과 그 예제를 나타낸다. 표 1의 우측 열은 문장 '한편 외국 자동차는 총 7만 2천 362대가 팔려 점유율이 39.8퍼센트로 떨어졌다.' 의 '372대가' 어절의 레이블을 결정할 때 사용되는 자질 예제이다.

표 1 한국어 의미역 결정 자질 예제

현재 어절에 포함된 형태소들의 어휘 및 품사 정보	362/SN, 가/JKS, SN, JKS, SN|NNB|JKS
현재 어절의 앞뒤 어절에 포함된 형태소들의 어휘, 품사 정보	2/SN, 팔리/VV, 천/NR, 어/EC, SN, VV, NR, EC, SN|NR, VV|EC,
서술어의 어휘 및 품사 정보	팔리/VV
서술어와 현재 어절의 위치 관계 및 거리 정보	PREV(위치 관계), DIST=1(텍스트 거리)

실험에 사용한 한국어 word embedding(단어 표현)은 word2vector[10]를 이용하여 구한 것을 사용하였다. feature embedding은 랜덤으로 초기화한 값을 사용하였고, 또한 Projection layer와 히든레이어에 Dropout[11]을 적용하였다. 성능 지표는 정확도와 재현율의 조화평균인 F1 지표를 사용하였으며, 본 논문에서 제시하고 있는 성능은 의미역 결정 문제의 논항 인식 및 분류(Argument Identification and Classification)에 해당한다. 평가의 단위는 어절 단위이며, Micro average를 사용한다.

표 2는 모델 별 한국어 의미역 결정 실험 결과이다. BiLSTM-CRFs 모델이 78.17의 성능을 보였고, Highway BiLSTM-CRFs 모델이 78.84로 BiLSTM-CRFs 모델보다 0.67% 더 높은 성능을 보였다. 앞선 두 모델의 성능차이를 통해 비선형 변환을 거치지 않은 입력 정보를 조절 하는 게이트가 성능 향상에 도움이 되는 것으로 유추할 수 있다. 추가적으로 각 모델에 히든레이어를 한 층 더 쌓아 실험을 진행하였는데, 실험 결과 Stacked BiLSTM-CRFs에서는 성능 향상이 있었으나 Stacked Highway BiLSTM-CRFs 모델에서는 경미한 성능 하락이 있었다. 또한 Highway Network와 유사하게 입력 정보를 가중치 연산 없이 히든레이어의 입력으로 사용하는 Residual Network[12]를 BiLSTM-CRFs에 적용하였으나, 실험 결과 가장 낮은 성능을 보였는데 Highway Network와 달리 가중치 연산이 없고, 입력 차원의 크기와 히든레이어의 크기가 같아야 하는 제한으로 인해 단어 정보 및 여러 자질 정보를 사용하는 의미역 결정에는 맞지 않는다고 생각한다.

표 2. 한국어 의미역 결정 실험 결과(AIC)

모델	F1
BiLSTM-CRFs (base)	78.17
Stacked BiLSTM-CRFs (2 layers)	78.57
Highway BiLSTM-CRFs	**78.84(+0.67)**
Stacked Highway BiLSTM-CRFs (2 layers)	78.77
Residual BiLSTM-CRFs	77.73(-0.44)

5. 결론

본 논문에서는 BiLSTM-CRFs 모델에 Highway Network를 적용하였고, 한국어 의미역 결정에서 기존 연구보다 좋은 성능을 얻었다. 이를 통해 비선형 변환을 거치지 않은 입력 정보가 한국어 의미역 결정 성능 향상에 도움이 됨을 알 수 있었다. 또한 Residual Network를 적용한 결과 입력 정보를 가중치 없이 히든레이어의 입력으로 사용하는 것이 한국어 의미역 결정의 성능 향상에는 도움이 되지 않는다는 것을 알 수 있었다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [2013-0-00131, (엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

참고문헌

[1] 배장성, 이창기, 임수종. 딥 러닝을 이용한 한국어 의미역 결정. 한국컴퓨터종합학술대회 논문집. 690-692. 2015.

[2] 배장성, 이창기. Stacked Bidirectional LSTM-CRFs를 이용한 한국어 의미역 결정, 정보과학회논문지, 제44권 제1호, 2017.

[3] Jie Zhou and Wei Xu, End-to-end learning of semantic role labeling using recurrent neural networks, Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 1127-1137, 2015.

[4] YAO, Kaisheng, et al. Spoken language understanding using long short-term memory neural

networks. In: Spoken Language Technology Workshop (SLT), IEEE. 189-194. 2014.
[5] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks", Advances in neural information processing systems, pp. 3104-3112, 2014.
[6] 박천음, 이창기. Bidirectional LSTM-CRF 모델을 이용한 멘션탐지, 한글 및 한국어 정보처리 학술대회, 2015.
[7] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. 2016. High- way long short-term memory rnns for distant speech recognition. In 2016 IEEE International Confer- ence on Acoustics, Speech and Signal Processing (ICASSP). pages 5755-5759.
[8] Luheng He, et al. Deep Semantic Role Labeling: What Works and What's Next. ACL 2016
[9] Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon, Korean Propbank [Online]. Available: http://catalog.ldc.upenn.edu/LDC2006T03.
[10] Tomas Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
[11] 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.
[12] He, Kaiming, et al. Deep residual learning for image recognition. arXiv:1512.03385 (2015).