JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 6, pp. 13-21, June 2023
https://doi.org/10.9708/jksci.2023.28.06.013

Analysis of dynamic LSTM network activation in LSTM English learner language model: Subject-Verb Number Agreement task

Euhee Kim*

*Professor, Dept. of Software Convergence, Shinhan University, Gyeonggi-do, Korea

[Abstract]

In this paper, we propose an approach to analyze the dynamic LSTM network activation in an LSTM English learner language model trained on a large corpus of English-speaking learners. The objective is to examine the relationship between network activation and performance on the Subject-Verb Number Agreement task. By employing a NA-task probing classifier and conducting ablation experiments, activation patterns are evaluated at each time step. The results reveal a strong link between network activation and the classifier's performance. The proposed model achieved 99.57% accuracy on the evaluation dataset for the NA-task, demonstrating its acquisition of correct grammar rules and accurate prediction ability. To analyze the influence of internal neurons on NA-task processing, specific LSTM neurons are removed and the model's performance is examined. Removing neuron 776 resulted in a more than 10% decrease in performance for plural subjects, while removing neuron 988 led to a 10% decrease in performance for singular subjects compared to the model before removal.

▸ **Key words**: probing classifier, LSTM language model, NLP, LSTM network, dynamic activation

[요 약]

본 논문에서는 대규모 영어학습자들의 말뭉치에서 학습된 LSTM 영어학습자 언어모델의 동적 LSTM 네트워크 활성을 분석하는 방법을 제안한다. NA-task 탐색 분류기(주어와 동사 간의 수일치 작업)를 전이 학습시키고 LSTM 네트워크에서 각 시점에서의 활성 및 LSTM 뉴런 유닛의 제거를 통해 성능을 평가하였다. 실험 결과는 LSTM 네트워크 활성이 NA-task 탐색 분류 성능에 영향을 미치는 것으로 나타났다. NA-task 탐색 분류 모델은 NA-task에 대해 평가 데이터셋에서 99.57%의 정확도를 달성하여 올바른 문법 규칙을 습득하고 정확한 예측 능력을 갖고 있음을 보여주었다. NA-task 처리에 대한 내부 LSTM 뉴런의 영향을 분석하기 위해 특정 LSTM 뉴런을 제거하고 모델의 성능을 평가 비교하였다. 뉴런 776을 제거한 결과, 복수 주어의 경우 성능이 10% 이상 감소하였으며, 뉴런 988을 제거한 결과, 단수 주어의 경우 성능이 제거 전 모델과 비교하여 10% 감소한 것으로 나타났다.

▸ **주제어**: 탐색 분류기, LSTM 언어모델, 자연어처리, LSTM 신경망, 동적 활성화

• First Author: Euhee Kim, Corresponding Author: Euhee Kim
*Euhee Kim (euhkim@shinhan.ac.kr), Dept. of Software Convergence, Shinhan University
• Received: 2023. 05. 15, Revised: 2023. 06. 05, Accepted: 2023. 06. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 몇 년 동안, 순환 신경망(RNN) 중 특히 장단기 메모리 신경망(Long Short-Term Memory, LSTM)을 이용한 언어모델은 다양한 자연어처리(NLP) 작업에서 성공적으로 적용되고 있다[1-2].

LSTM은 NLP 분야에서 텍스트 문장을 처리하는 데 사용되며, 데이터 학습 과정에서 LSTM이 언어의 문법적 특성을 포착하고 있는지를 확인하기 위해 탐색 분류기(probing classifier) 또는 신경망 언어모델을 사용하여 NLP 작업의 언어 성능을 평가한다[3-8]

탐색 분류기는 사전 학습된(pre-trained) LSTM 언어모델을 레이블이 지정된 데이터에 대한 추가 학습 없이, 특정 언어 정보를 추출하는 데 사용한다. 이러한 방식은 언어모델의 내부 표상(embedding representation)을 분석하고 해석하는 것뿐만 아니라, 그러한 정보를 필요로 하는 다운스트림(downstream) NLP 작업에 유용한 접근법으로 이용된다[9].

신경망 언어모델을 이용하는 탐색 분류기는 다른 방식으로 NLP 작업의 언어 성능을 평가한다. 예를 들어, 영어 문장의 문법적인 정확성을 판단하기 위해서는 탐색 분류기가 범용의 신경망 언어모델에 의존하지 않고 더 구체적으로 설계될 수 있다. 또한, 해석 가능한 결과를 제공하여 연구자들이 신경망 언어모델이 포착하지 못하는 특정 언어적 특징을 파악할 수 있도록 도와주며, 비교적 저렴한 비용으로 많은 양의 데이터에 대해 훈련할 수 있다. 이러한 방식으로 탐색 분류기는 NLP 모델의 언어 성능을 평가하고, 이를 개선하는 데 중요한 역할을 한다.

탐색 분류기는 다양한 언어모델의 다른 NLP 작업으로 전이 학습(transfer learning) 가능성을 평가하는 방법을 제공한다. 그러나 이러한 NLP 작업이 어떻게 놀라운 성능을 이루는지에 대한 신경망의 내부 활성화에 대한 분석 연구는 많지가 않다.

본 논문에서는 탐색 분류기를 활용하여, 사전 훈련된 학습자 LSTM 신경망 언어모델에서 동적 LSTM의 활성을 추적하여, 영어 문장의 주어 동사 수일치 작업(Number agreement task; 이하 NA-task) 처리 방식을 분석하는 방법을 제안한다. NA-task은 문법 규칙의 일관성을 평가하기 위한 작업 중 하나로, 단수와 복수의 형태를 가진 명사와 그에 맞는 동사의 일치를 평가한다.

NA-task은 영어 학습 교육 분야에서 매우 중요하다. 이 작업을 통해, 영어학습자들이 명사와 동사 간의 일관성을 이해하고 있는지, 올바른 문법 규칙을 적용할 수 있는지, 그리고 언어적 지식의 수준을 평가할 수 있다.

영어 문장에 대한 품사 태깅, 표제어 추출과 의존 구문 분석 등 NLP 작업에 대한 해외 및 국내 연구가 많이 있지만 본 논문에서 제안한 영어 학습자들이 사용하는 영어 교재 코퍼스로 학습한 LSTM 언어모델을 이용하여 영어 문장의 NA-task에 대한 탐색 분류기 접근 방식을 적용한 연구는 거의 없는 것으로 안다.

본 논문에서는 초거대(large-scale) BERT나 GPT 언어모델과는 달리, 비교적 작은 규모의 영어 코퍼스로 훈련된 LSTM 신경망 언어모델이 NA-task 처리의 문법을 학습하는 과정에서 LSTM 네트워크의 동적인 활성이 NA-task 탐색 분류 성능에 미치는 영향을 분석한다[10-11].

본 논문의 구성은 다음과 같다. 2장에서는 LSTM 신경망 언어모델을 이용한 탐색 분류기 관련 연구들에 대해 설명하고, 3장에서는 제안 모델을 설명한다. 4장에서는 제안한 모델의 구현 방법에 대해 기술한다. 5장과 6장에서는 실험 결과를 기술하고, 결론을 맺는다.

II. Related works

1. Probing classifier

탐색 분류기는 자연어 처리의 심층 신경망 모델을 해석하고 분석하는 데 있어서 주요한 방법론 중 하나이다. 탐색 분류기의 동작 원리는 어떤 언어학적인 특성을 모델의 표현으로부터 예측하기 위해 분류기를 훈련시킨다.

탐색 분류기는 심층 신경망 모델의 내부 동작을 해석하고 이해하는 데 있어서 다음과 같은 장점들을 가지고 있다. 첫째, 모델이 어떤 특성을 학습하고 있는지 분류기를 통해 확인할 수 있다. 둘째, 모델이 어떤 특성을 기반으로 예측을 수행하는지 설명할 수 있다. 따라서 모델이 잘못된 예측을 하는 경우, 분류기를 통해 그 이유를 파악하고 모델을 개선할 수 있다. 또한, 모델이 특정 데이터 셋에만 과적합(overfitting)되는 것을 방지할 수 있다. 따라서 탐색 분류기는 자연어 처리 모델의 해석과 분석에 있어서 유용한 도구이다[9].

본 연구에서는 학습자 LSTM 언어모델을 이용한 탐색 분류기의 아키텍처는 예측 작업을 수행하기 위해 입력 임베딩 계층(input embedding layer), LSTM 계층(LSTM layer), 풀링 계층(pooling layer), 탐색 계층(probing layer), 학습(training) 단계로 구성하였다. 이 탐색 분류기를 이용하여 LSTM 언어모델에 대한 NA-task 모델을 설계하였다.

2. LSTM language model

LSTM 신경망 언어모델은 문장 생성에 적합한 모델이다. LSTM 언어모델은 기존의 피드포워드 신경망과 달리 입력 데이터를 순차적으로 처리하는 대신 LSTM 메모리 셀을 사용하여 정보를 저장하며, 이전 셀 상태 및 입력에 기반을 두어 정보를 선택적으로 잊거나 기억한다[1-3].

LSTM 언어모델은 자연어 처리에 매우 유용하다. 이는 문장이나 단락과 같은 자연어에서 발생하는 장기적인 의존성을 효과적으로 포착할 수 있기 때문이다. 이 모델은 단어나 문자의 순서를 고려하고, 이전 순서를 기반으로 다음 단어나 문자의 가능성을 예측하는 방식으로 작동한다. 대규모 텍스트 데이터 말뭉치에서 학습되며, 입력 및 출력 계층을 조정하여 NLP 작업에 대해 언어모델을 미세조정(fine-tuning)할 수 있다.

특히, Gluordava et al.가 제안한 LSTM 언어모델은 NLP 작업 관련 광범위한 응용 분야에서 사용되어 왔다. 이 언어 모델은 단방향의 2개 LSTM 계층으로 구성되며, 각 계층은 입력 단어의 임베딩 벡터를 입력으로 받아들이고 은닉 상태 벡터를 출력한다. 그런 다음 한 LSTM 계층의 은닉 상태 벡터가 다음 LSTM 계층에 입력으로 공급되어진다. 그런 다음 최종 은닉 상태가 소프트맥스 계층에 공급되어 다음 단어에 의한 확률 분포를 생성한다. 이 모델은 순차적으로 예측된 다음 단어 사이의 교차 엔트로피 손실을 최소화하기 위해 역전파 시간을 사용하여 훈련된다. 학습 데이터는 위키피디어 기사, 뉴스 기사 및 웹페이지를 포함한 다양한 텍스트로 구성되었고 또한 학습 데이터의 크기를 늘리고 언어 모델의 성능을 향상시키기 위해 데이터 증강 알고리즘을 사용하였다. 이 모델은 Penn Treebank 및 WikiText-103 데이터 세트를 포함한 표준 벤치마크를 사용하여 모델의 성능을 평가했으며 이러한 작업에서 다른 최첨단 언어 모델보다 성능이 우수함을 발표하였다[2].

본 연구에서 사용한 LSTM 언어모델은 Gulordava et al. 에 의해 개발된 언어모델의 아키텍처를 기반으로 영어 학습자의 영어 교재와 데이터 증강 기법을 사용하여 추가한 학습데이터로 훈련하였다[14-17].

3. LSTM memory update and output rules

LSTM 뉴런은 RNN에서 발생할 수 있는 기울기 소실 문제로 인해 자연어와 같이 긴 문장을 학습하는 것이 어려워지는 문제를 해결하기 위해 고안되었다[1].

언어모델의 학습을 위해 LSTM 신경망이 사용될 때, 신경망의 가중치는 주어진 입력에 대한 예측 출력과 실제 출력 간의 차이를 기반으로 업데이트된다. 이 학습과정은 여러 번 반복되며, 신경망의 예측이 충분히 정확해질 때까지 진행된다. 학습 과정에서 LSTM 뉴런의 내부 상태는 학습 데이터의 장기적 의존성을 포착하는 방법을 학습하면서 점진적으로 변화한다.

LSTM 뉴런 내부는 정보 흐름을 제어하는 세 가지 유형의 게이트(gate)로 구성된다. 입력 게이트는 셀(cell) 상태에 추가할 새 정보의 양을 결정하며, 학습 중에는 현재 입력(x_t), 이전 장기메모리 셀 상태(C_{t-1}), 그리고 단기메모리 은닉 상태(h_{t-1})를 기반으로 셀 상태를 선택적으로 업데이트(C_t)하는 방법을 학습한다. 망각 게이트는 셀 상태에서 제거할 정보의 양을 결정하여 LSTM 메모리가 관련 없는 정보를 폐기하고 당면한 작업에 대한 관련 정보만 유지할 수 있도록 한다. 출력 게이트는 현재 시간 단계를 예측하기 위해 사용할 셀 상태의 양을 결정하며, 셀 상태에서 관련 정보를 선택적으로 사용하여 정확한 예측을 할 수 있도록 학습한다. 이 과정은 LSTM 셀 상태의 점진적 변화로 이어져 네트워크가 언어모델링 등 광범위한 언어 작업을 학습하고 정확한 예측을 수행할 수 있게 된다.

수식 (1)과 (2)는 셀 업데이트 규칙을 나타낸다. 여기서 각 게이트 스칼라 값인 f_t, i_t, o_t는 학습 과정에서 활성화 함수(tanh)에 의해 계산된다. $\widetilde{C_t}$는 업데이트할 현재 시간 단계의 셀 상태를 나타낸다. 또한, h_t는 현재 시간 단계의 은닉 상태를 나타낸다. 이러한 규칙에 따라 LSTM 메모리는 장기적인 의존성을 갖는 데이터를 모델링할 수 있다.

$$C_t = f_t \circ \ C_{t-1} + i_t \circ \ \widetilde{C_t} \qquad (1)$$

$$h_t = o_t \circ \ \tanh(C_t) \qquad (2)$$

4. English-based Number-Agreement task

주어와 동사간의 수일치(number-agreement task; 이하 NA-task) 작업은 영어 문장에서 문법 규칙의 일관성을 평가하기 위한 작업 중 하나이다. 이 작업에서는 단수와 복수의 형태를 가진 명사와 그에 맞는 동사의 일치를 평가한다.

NA-task는 다음과 같이 수행된다. 먼저, 두 개의 문장이 제시된다. 이 두 문장은 구문적으로 유사하지만, 명사와 동사의 복수와 단수 형태가 다르다. 그런 다음, LSTM 언어모델을 이용한 탐색 분류기는 두 문장 중에서 어느 문장이 올바른 문장인지 선택하게 된다. 예를 들어, “The boy greets”와 “The boy greet” 두 문장이 제시될 때, “The boy greets”가 올바른 문장이므로 NA-task 모델은 이 문장을 예측한다.

일반적으로 NA-task는 문법적으로 틀린 다수의 문장들을 제시하고, 주어와 동사 사이의 일치 여부를 정확하게 판단하여 오류를 수정하는 작업이기 때문에 영어 문장을 이해하기 어렵게 만들 수 있어, 언어 모델의 영어 문장 이해 능력을 측정하는 데에 사용된다. 본 연구에서는 NA-task에서 LSTM 메모리의 셀 상태를 추적하기 위해 Lizen et al.이 구축한 데이터 셋을 사용하여 LSTM 언어 모델 기반을 둔 탐색 분류기인 NA-task 모델을 구현하였다.

Table 1은 NA-task에서 사용한 Linzen et al. 데이터 셋의 예문 일부를 정리한 표이다. 데이터 셋은 20개의 주어/목적어 명사, 15개의 동사, 10개의 부사, 5개의 전치사, 10개의 고유명사, 10개의 위치 명사 풀에서 단어를 무작위로 선택하여 4800 개 문장들로 구성된다. 각 문장의 주어와 동사의 수를 변경하여 일치 및 불일치 조건을 적용하여 정문(correct)과 비문(wrong)을 구분한다. 예를 들어, 문장의 주어와 동사가 동시에 단수로 일치하는 정문 문장은 SS(singular-singular) 와 correct, 불일치하는 비문 문장은 SP(singular-plural)와 wrong 조건으로 표시된다[7, 12].

Table 1. Dataset for NA-task

Condition	Sentences	C/W
SS	The **boy** near the cars **greets**	correct
SP	The **boy** near the cars **greet**	wrong
PS	The **boys** near the car **greets**	wrong
PP	The **boys** near the car **greet**	correct

III. The Proposed NA-task Classifier

LSTM 언어모델을 기반으로 한 탐색 분류기 모델로 설계한 NA-task 모델은 주어와 동사의 문법적 수일치를 예측하는 자연어 처리 모델이다. 이 모델은 LSTM 메모리 내부 상태를 사용하여 문장 내의 단어들 간의 문맥과 의존성을 추적하고, 입력 단어 시퀀스와 내부 상태를 기반으로 문법적 수 일치에 대한 예측을 수행한다.

Fig. 1은 NA-task 모델의 구성도이다. Table 1의 Linzen et al. 데이터 셋을 이용하여 NA-task 모델 학습에 필요한 데이터 셋을 생성하고(generate), 사전 학습된 LSTM 언어모델을 미세조정((load)하는 구조이다.

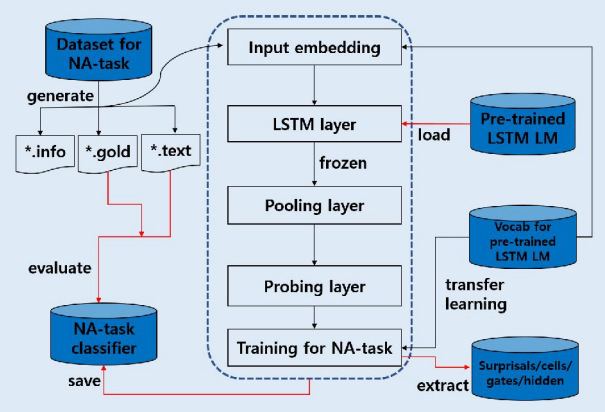

Fig. 1. Architecture of NA-task model for the LSTM LM

NA-task 모델은 전이학습(transfer learning)을 통해 사전 학습된 LSTM 언어모델과 단어사전(vocab), 3개의 파일(*.text, *.gold, *.info)로 NA-task 모델을 학습하고 성능을 평가한다.

우선, 입력 임베딩 계층(input embedding layer)에서는 입력 문장의 각 단어를 임베딩 벡터로 표현하여 LSTM 계층(LSTM layer)에 입력한다. LSTM 계층은 각 시점에서 NA-task의 문장에서 문맥 정보를 포착하며, LSTM 뉴런의 셀과 은닉 상태와 게이트 정보를 출력한다. LSTM 계층 위 풀링 계층(pooling layer)에서는 global max pooling을 적용하여 은닉 벡터의 길이를 고정시키고 시간 단계별 출력을 가져와서 각 차원별로 최대값을 선택하여 문맥 문장에서 가장 중요한 정보를 추출한다. 풀링 계층 위 탐색 계층(probing layer)은 feedforward, fully connected 계층, log-softmax 출력 계층으로 구성되며, 활성화 함수 sigmoid를 이용하여 NA-task 관련 주요 정보 속성을 추출하고 추적한다. 이 과정에서 LSTM 계층은 추가 학습을 하지 않고(frozen), 탐색 계층에서 업데이트가 수행된다. NA-task 모델의 성능은 *.text 파일과 *.gold 파일을 사용하여 평가하며, 수일치 문장의 예측 분류결과를 LSTM 메모리 내부 구조의 점진적 상태 변화와 관련하여 해석할 수 있다.

IV. Experiments

Fig. 1의 영어학습자 LSTM 언어모델 기반 NA-task 모델은 Pytorch로 구현하였으며, 사용한 하드웨어는 Table 2와 같다.

Table 2. Hardware configuration

Name	Version
GPU	RTX A5000
CPU	i9-10980XE
RAM	32 GB
HDD	2TB
SSD	1TB

1. Data Augmentation Using Bert model

초거대 BERT 나 GPT 언어모델은 대규모 코퍼스로 학습되어 다양한 NLP 작업에서 최고 성능을 보이며, 입력 문장의 문맥 임베딩 학습을 통해 문법적으로 올바른 문장 생성을 생성할 수 있다.

본 연구에서는 BERT 또는 GPT 모델보다 상대적으로 규모가 작은 영어학습자 LSTM 언어모델을 이용하여 NA-task 모델을 구현하고자 한다. NA-task 모델의 성능은 영어학습자 LSTM 언어모델의 성능과 밀접한 상관이 있기 때문에 V. Kumar et al.의 데이터 증강 알고리즘을 적용하여 기존의 영어학습자 코퍼스(L2)를 추가로 증강시켜 L2 LSTM 언어모델의 성능을 개선하였다[13].

Table 3. Data Augmentation for L2 Corpus

Algorithm: Data augmentation for L2 text

1. Input:
   Dataset D_{train}, Pre-trained model BERT
2. Preprocess D_{train}:
- Converts all text data to lowercase
- Removes unnecessary symbols
- Removes punctuation
- Tokenize sentences
- Transforms tokenized sentences to fit the input format of the BERT model.
3. Fine-tune BERT using D_{train}:
$D_{synthetic} \leftarrow []$
for each sentence in preprocessed D_{train}:
  mask the selected words
  synthesize sentences using BERT
  $D_{synthetic} \leftarrow D_{synthetic} \cup$ sentences

Table 3의 BERT 모델을 이용한 데이터 증강 학습에서 사용한 L2 학습 코퍼스는 한국에서 2016-2018년 출판된 EBS-CS 영어교과서, 2001년 출판된 11개 중학교와 12개 고등학교 영어교과서 그리고 2009년에 출판된 19개 중학교와 12개 고등학교 영어교과서를 포함한다.

V. Kumar et al.의 데이터 증강 알고리즘은 L2 학습 코퍼스 문장에 마스킹(masking) 기법을 적용하여 문장의 다양성을 높여 새로운 문장을 생성한다. 마스킹 기법은 문장을 전 처리(preprocess)하여 입력 문장(Dtrain)에서 무작위로 단어를 선택하여 마스킹하고, 해당 단어를 예측하도록 BERT 언어모델에 입력하여 추가 학습을 시켜 문장들을 예측한다. 예측 문장 중 무작위로 한 개를 선택하여 마스킹 처리된 단어를 대체하여 새로운 문장(synthetic)을 증강한다.

본 연구에서는 영어 지식수준이 다른 L2 LSTM 언어모델을 모델링하기 위해 증강한 L2 코퍼스를 크기 별로 2개로 나누어 2개 버전의 L2 LSTM 언어모델(LSTM-small 및 LSTM-large)을 구현하는 데 사용하였다. 실험에서 사용한 L2 코퍼스는 Table 4와 같다.

본 실험에서는 L2 LSTM-small 언어모델의 NA-task 모델을 베이스라인으로 하여 L2 LSTM-large 언어모델의 NA-task 모델과 성능을 비교하였다.

Table 4. L2 Corpus for L2 LSTM language model

L2 LSTM LM	L2 Corpus
L2 LSTM-large	102.5M tokens
L2 LSTM-small	52.5M tokens

2. Metrics

본 연구에서는 L2 LSTM 언어모델의 성능을 측정하는 지표로서 PPL을 사용하였고, NA-task 모델에 대한 성능을 평가하는데 정확도(accuracy) 측정 지표를 사용하였다.

NA-task 모델은 베이스라인 모델로서 L2-small 모델과 L2-large 모델을 구현하여 성능을 평가 비교하였다. L2-large 모델은 영어학습자 코퍼스 1억 단어로 구현한 LSTM-large 언어모델 기반을 둔 NA-task 모델을 표기하며, L2-small 모델은 5000천만 단어로 구현한 LSTM-small 언어모델 기반을 둔 NA-task 모델을 의미한다.

PPL 지표는 언어모델의 학습 및 평가에서 사용되며, 모델이 주어진 문장의 다음 단어를 예측하는데 얼마나 힘든지를 나타내는 수치이다. 모델이 예측하려는 시퀀스에서 각 단어가 나타날 확률의 역수를 계산하여 얻는다.

정확도 지표는 NA-task 모델이 앞의 단어의 맥락을 고려하여 테스트 문장(*.text)에서 주어와 동사 간의 수일치를 올바르게 평가한 비율을 나타낸다. 구문적으로 유사하지만 주어와 동사의 복수와 단수 형태가 다른 두 개의 문장을 모델에 제시하고, 모델의 출력 확률을 비교했다 ($surprisal(V_C) < surprisal(V_W)$). 정확하게 선택한 동사 형태의 확률이 잘못된 동사 형태보다 높을 경우 결과 점수는 1, 그렇지 않으면 0으로 처리하였다. 이러한 결과들의 합계를 전체 테스트 문장 수로 나눈 비율로 측정하였다. 정확도의 수식은 다음 (3)과 같다.

$$Accuracy = \frac{\sum[surprisal(V_C) < surprisal(V_W)]}{No.\ of\ \ sentences} \quad (3)$$

3. The neuron ablation mechanism of probing LSTM representations for a NA-task

본 연구에서는 NA-task 작업에 대한 LSTM 언어모델의 내부 LSTM 뉴런을 탐색하기 위해 특정 뉴런을 제거하는 기법을 적용하여 성능을 분석하였다.

뉴런 제거 기법은 언어모델의 특정 뉴런을 제거한 후, 모델 성능을 비교하는 데에 활용된다. 이를 통해, 모델 내의 특정 뉴런이 언어 처리에서 어떤 역할을 하는지 이해할 수 있다.

Fig. 1의 LSTM 계층의 뉴런이 NA-task 작업 성능에 어떠한 영향을 미치는지 확인하고자, 특정 뉴런을 제거한 NA-task 모델과 전체 NA_task 모델 간의 성능을 비교할 때, 특정 뉴런의 모든 입력 또는 출력 뉴런을 제거한 후, 그 모델의 성능을 수식 (3)의 정확도를 이용하여 측정하였다.

특정 뉴런의 영향력을 분석하기 위해, 특정 뉴런을 제거할 때 모델 성능에 어떤 영향을 미칠지 예측할 수 있도록 뉴런의 입력과 출력을 분석하고, 뉴런이 제거될 때, 먼저 NA-task 작업에서 모델 성능을 평가하기 전에 모든 다른 가중치를 0으로 설정하였다.

4. Hyperparameters

제안한 NA-task 모델에서 활용한 사전학습 L2 LSTM 언어모델 구현에 사용했던 주요 하이퍼파라미터는 Table 5와 같다.

LSTM-small과 LSTM-large 언어모델 학습은 Table 5에 표기된 조건을 모든 태스크에 동일하게 적용하였다.

Table 5. A summary of model hyperparameter

L2 LSTM	Hyperparameter	
pretrained model	type of recurrent net	LSTM
	number of hidden units per layer	650
	size of word embedding	650
	vocabulary size	69,577
	number of layers	2
	learning rate	$1e^{-3}$
	batch size	128
	sequence length	35
	dropout	0.2
	epoch	40
	input dim	768
	optimizer	Adam

V. Results

본 장에서는 Linzen et al.에서 제안한 NA-task 데이터셋을 이용하여 영어학습자 LSTM-small, LSTM-large 언어모델 기반 NA-task 모델들로 학습한 실험 결과를 비교 기술한다.

Table 6은 구현한 모델들의 성능 결과를 정리하였다. LSTM 언어모델 기반 NA-task 모델의 성능 평가 지표 PPL과 정확도는 학습 데이터셋의 크기에 따라 L2-small 모델은 각각 110.20, 89.3%, L2-large 모델은 51.01, 99.57% 결과가 나왔다.

Table 6. A summary of performance of the LSTM languages models and NA-task models

model / metric	LSTM-small	LSTM-large
PPL	110.20	51.01
model / metric	L2-small	L2-large
Accuracy(%)	89.3	99.57

Table 6의 결과는 L2-large 모델이 L2-small 모델보다 NA-task 작업을 상대적으로 정확하게 예측할 수 있음을 보여준다. 이는 LSTM-small 언어모델의 성능이 LSTM-large 언어모델의 성능보다 상대적으로 낮기 때문에 L2-small 모델보다 L2-large 모델이 올바른 문법 규칙을 습득하고 올바르게 예측할 수 있는 능력을 갖고 있다고 분석할 수 있다.

Table 7은 NA-task 모델의 LSTM 언어모델의 내부 뉴런의 영향력을 분석하기 위해 특정 뉴런 제거 기법을 적용하여 모델 성능을 분석한 결과이다. 여기서 Ablated neuron unit 항목은 제거한 특정 뉴런, Full 항목은 제거한 뉴런 없음, C 항목은 주어와 동사 간의 형태로서, SS은 단수-단수 형태, SP은 단수-복수 형태, PS은 복수-단수 형태, PP은 복수-복수 형태를 표시하며, 그리고 10% 미만의 성능 감소는 '-'로 표시한다.

Table 5의 LSTM 언어모델 구조는 은닉 계층의 차원이 650이고 LSTM 계층을 2개 사용하므로 첫 번째 계층의 뉴런들을 표시할 때는 왼쪽부터 1에서 650까지 순번을 매기고, 두 번째 계층에 속하는 뉴런들을 표시할 때는 651부터 1300까지 순번을 지정하였다. Table 7의 분석결과는 두 번째 계층에 속하는 776 뉴런 유닛을 제거했을 때 L2-large 모델의 성능은 주어가 복수인 경우 제거 전 모델의 성능보다 10% 이상 감소했다. 또한, 두 번째 계층에 속하는 988 뉴런 유닛을 제거했을 때 L2-large모델의 성능은 주어가 단수인 경우 제거 전 모델의 성능보다 10% 감소한 결과가 나왔다. L2-small 모델은 특정 뉴런 제거 전과 제거 후의 성능의 차이는 10% 미만으로 나왔다.

Table 7. The accuracy(%) of performance of the LSTM languages model on this NA-tasks with 4 contrasts.

model	C	Ablated neuron unit (verb)		Full
		776 (plural)	988 (singular)	
L2-large	SS	-	-	99.3
	SP	-	54.1	99.7
	PS	48.0	-	99.2
	PP	78.3	-	99.9
L2-small	SS	-	-	88
	SP	-	-	89
	PS	-	-	87.4
	PP	-	-	89.6

Table 7의 결과로부터 사전학습 LSTM 언어모델의 특정 뉴런 단위 776과 988을 각각 제거한 후, 제거 전의 모델 성능 결과와 비교해 볼 때, L2-large 모델이 L2-small 모델보다 모델 내의 특정 뉴런이 NA-task 작업처리에서 중요한 역할을 하고 있음을 확인할 수 있다.

Fig. 2는 L2-large 모델을 이용하여 주어와 동사 사이의 전치사구가 있는 문장을 처리하는 동안 LSTM 내부 뉴런의 셀과 게이트에 대한 동작을 탐색한 결과를 그린 도식도이다. 위 패널 그래프는 LSTM 계층의 뉴런 776 탐색에 대한 그래프이고, 아래 패널 그래프는 뉴런 988에 대한 그래프이다. 파란색 실선은 Table 7의 C항목이 PP형태, 빨간색 실선은 SS형태, 파란색 점선은 PS형태, 빨간색 점선은 SP형태를 표시한다.

Fig. 2는 위 패널과 아래 패널 그래프의 셀 C_t, $\widetilde{C}_t$ 그리고 입력 게이트 i_t에 대한 PP와 PS의 빨간색 선과 SS와 SP의 파란색 선을 각각 비교해 볼 때 복수동사는 단수동사 뉴런과 유사한 패턴을 갖고 있음을 읽을 수 있다.

Fig. 2의 그래프 결과에 의하면, L2-large 모델이 주어와 동사 간의 문법적 수를 식별하는 방식이 주어와 동사 사이의 전치사구와 같은 문법적인 특징에 의해 영향을 받는 것임을 알 수 있다. 예를 들어, 주어 명사(boy 또는 boys)와 동사(greets 또는 greet) 시점에서 곡선의 변동이 크게 나타남을 확인할 수 가 있다.

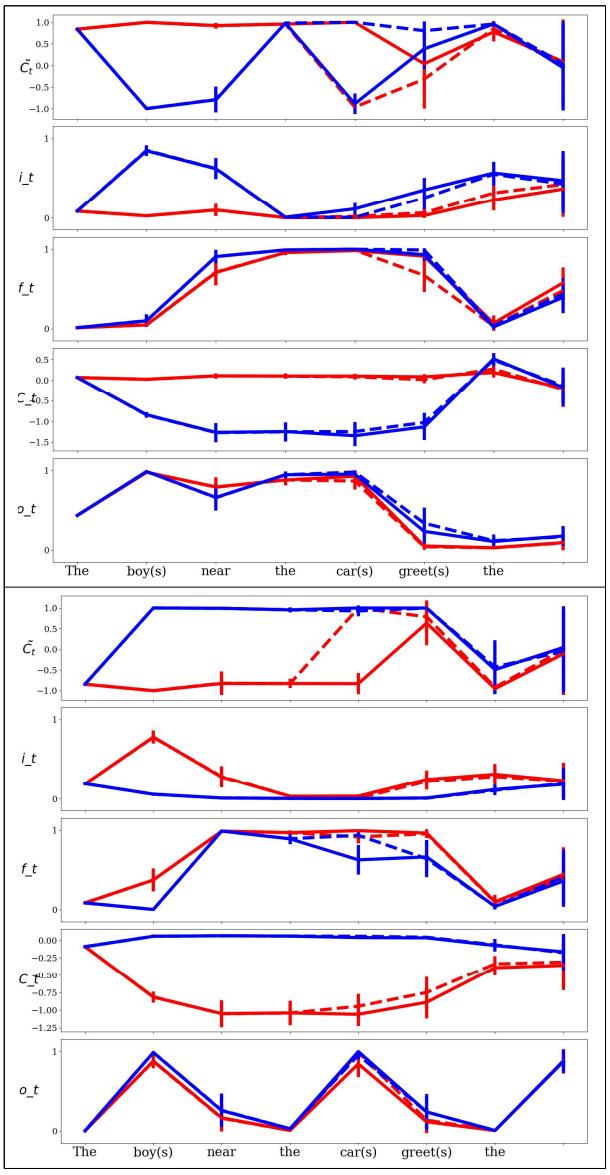

Fig. 2. L2-large model : Cell units and gate activations during processing of sentences between subject and plural verb(776;top panel) and singular verb(988;bottom panel)

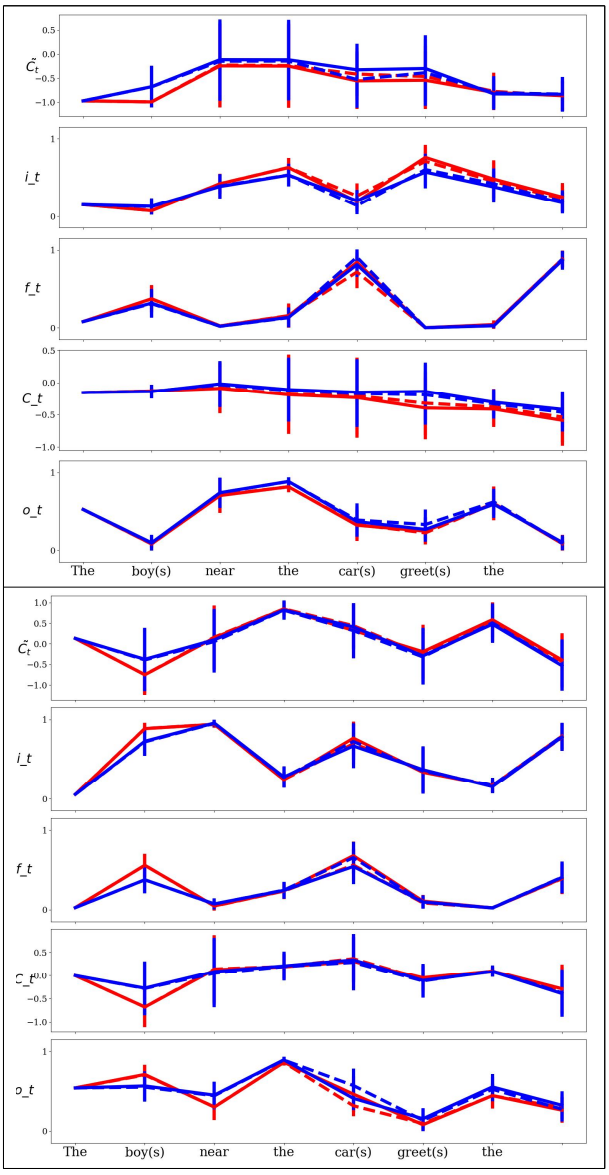

Fig. 3. L2-small model : Cell units and gate activations during processing of sentences between subject and plural verb(776;top) and singular verb(988;bottom)

Fig. 3은 L2-small 모델을 이용하여 주어와 동사 사이의 전치사구가 있는 문장을 처리하는 동안 LSTM 내부 뉴런의 셀과 게이트에 대한 동작을 탐색한 결과를 그린 도식도이다.

L2-large 모델과 달리, L2-small 모델의 LSTM 계층의 뉴런 776과 988은 주어와 동사 사이에 전치사구가 있는 문장을 처리할 때 셀 C_t과 $\widetilde{C}_t$ 그리고 입력 게이트 i_t의 PP와 PS의 빨간색 선과 SS와 SP의 파란색 선을 각각 비교해 볼 때 곡선 간의 변동 차이가 상대적으로 크지 않음을 확인할 수 있다. 이를 통해 L2-small 모델은 L2-large 모델보다 올바른 문법 규칙을 습득하고 올바르게 예측할 수 있는 성능은 낮지만, L2-small 언어모델이 주어와 동사 간의 문법적 수를 식별하는 방식이 주어와 동사 사이의 전치사구와 같은 문법적인 특징에 의해 영향을 받는 것임을 알 수 있다.

VI. Conclusions

본 논문에서는 대규모 영어학습자들의 말뭉치에서 학습된 LSTM 영어학습자 언어모델의 동적 LSTM 네트워크 활성을 분석하였다.

NA-task 모델은 LSTM-large와 LSTM-small 언어모델을 미세 조정하여 전이학습을 통해 LSTM 계층에서 각 시점에서 뉴런 셀과 게이트 정보를 추출하였다. 추출한 정보를 사용하여 NA-task 작업에 대한 제안된 모델의 성능을 비교 및 분석하였다.

NA-task 모델은 Linzen et al. 평가 데이터셋을 대상으로 L2-large 모델의 성능은 99.57% 정확도가 나왔다. L2-large 모델이 NA-task에 대한 올바른 문법 규칙을 습득하고 올바르게 예측할 수 있는 능력을 보여준다.

NA-task 처리에 대한 LSTM 언어모델의 내부 뉴런의 영향력을 분석하기 위해서 특정 LSTM 뉴런을 제거하여 모델의 성능을 분석하였다. 뉴런 776을 제거한 L2-large 모델의 성능은 주어가 복수인 경우 제거 전 모델의 성능보다 10% 이상 감소했으며, 뉴런 988을 제거한 L2-large 모델의 성능은 주어가 단수인 경우 제거 전 모델의 성능보다 10% 감소한 결과가 나왔다. 따라서 L2-large 모델은 특정 뉴런이 NA-task 작업처리에서 중요한 역할을 하고 있음을 의미한다.

본 연구 결과는 LSTM 네트워크 활성이 NA-task 탐색 분류 성능과 밀접한 관련이 있음을 보여주었다. 더불어 LSTM 네트워크의 동적인 활성이 NA-task 탐색 분류 성능에 영향을 미치는 것으로 나타났다. 이러한 결과는 LSTM 네트워크가 자연어 처리에 미치는 영향을 이해하는데 도움이 되며, LSTM 네트워크의 설계 및 개발에 기여할 수 있다.

ACKNOWLEDGEMENT

This work was supported by 2022 Shinhan Univ. Research Grant.

REFERENCES

[1] S. Hochreiter et al., "Long short-term memory," Neural Computation, pp. 1735-1780, Nov 1997.

[2] K. Gulordava et al., "Colorless green recurrent networks dream hierarchically," Proceedings of NAACL, pp. 1195-1205, Mar 2018.

[3] S. Chowdhury et al., "RNN simulations of grammaticality judgementss on long-distance dependencies," Proceedings of COLING, pp. 113-144, Aug 2018.

[4] M. Giulianelli et al., "Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information," Proceedings of the EMNLP BlackboxNLP Workshop, pp. 240-248, Nov 2018.

[5] D. Hupkes et al., "Visualization and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure," Journal of Artificial Intelligence Research, pp. 907-926, Apr 2018.

[6] T. Linzen et al., "Syntactic Data Augmentation Increases Robustness to Inference Heuristics," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 427-432, Apr 2020.

[7] T. Linzen et al., "The reliability of acceptability judgements across languages," Glossa: a journal of general lingistics, pp. 1-25, Sep 2018.

[8] T. Linzen et al., "Syntactic Structure from Deep Learning," Annual Review of Linguistics, pp. 195-212, Jan 2020.

[9] Y. Belinkov, "Probing Classifiers: Promises, Shortcomings, and Advances," Computational Linguistics, pp. 207-219, Apr 2022.

[10] J. Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," May 2019. https://arxiv.org/abs/1810.04805v2

[11] A. Radford et al., "Improving Language Understanding by Generative pre-Training," Aug 2018. https://arxiv.org/abs/1810.04805v2

[12] T. Linzen et al., "Assessing the ability of LSTMs to learn syntax-sensitive dependences," transactions of the Association for Computational Linguistics, pp. 521-535, Nov 2016. DOI:10.1162/tacl_a_00115

[13] V. Kumar et al., "Data Augmentation using pre-trained transformer models," AACL, pp. 195-212, Jan 2021

[14] K. Euhee, "Probing Sentence Embeddings in L2 Learners' LSTM Neural Language Models Using Adaptation Learning," Journal of Korean Institute of Information Scientists and Engineers, Vol. 27, No. 3, pp. 13-23, Mar 2022.

[15] K. Euhee et al., "A Neural Language Model as a Language Learner: Focusing on Subject-Verb Agreement," Language & Information Society, Vol. 44, No. 3, pp. 165-191, Mar 2021. DOI: 10.29211/soli.2021.44..006

[16] K. Euhee, "The Ability of L2 LSTM Language Models to Learn the Filler-Gap Dependency," Journal of Korean Institute of Information Scientists and Engineers, Vol. 25, No. 11, pp. 27-40, Nov 2020. DOI:10.9708/jksci.2022.27.03.013

[17] K. Euhee et al., "L2ers' predictions of syntactic structure and reaction times during sentence processing," Language & Information Society, Vol. 37, pp. 189-218, Sep 2020. DOI: 10.29211/soli.2021.44..006

Authors

Euhee Kim received the M.S. degrees in Computer Engineering from Dongguk University, Korea, in 2002 and Ph.D. degrees in Mathematics from The University of Connecticut, U.S.A in 1995.

Euhee Kim is currently a Professor in the Department of Software Convergence at Shinhan University. She is interested in AI, NLP and Big Data computing.