순환 신경망에서 LSTM 블록을 사용한 영어와 한국어의 시편 생성기 비교

에런 스노버거 · 이충호*

한밭대학교

Psalm Text Generator Comparison Between English and Korean Using LSTM Blocks in a Recurrent Neural Network

Aaron Daniel Snowberger · Choong Ho Lee*

Hanbat National University

E-mail : aaron@edu.hanbat.ac.kr / chlee@hanbat.ac.kr

요 약

최근 몇 년 동안 LSTM 블록이 있는 RNN 네트워크는 순차적 데이터를 처리하는 기계 학습 작업에 광범위하게 사용되어왔다. 이러한 네트워크는 주어진 시퀀스에서 가능성이 다음으로 가장 높은 단어를 기존 신경망보다 더 정확하게 예측할 수 있기 때문에 순차적 언어 처리 작업에서 특히 우수한 것으로 입증되었다. 이 연구는 영어와 한국어로 된 150개의 성경 시편에 대한 세 가지 다른 번역에 대해 RNN/LSTM 신경망을 훈련하였다. 그런 다음 결과 모델에 입력 단어와 길이 번호를 제공하여 훈련 중에 인식한 패턴을 기반으로 원하는 길이의 새 시편을 자동으로 생성하였다. 영어 텍스트와 한국어 텍스트에 대한 네트워크 훈련 결과를 상호 비교하고 개선할 점을 기술한다.

ABSTRACT

In recent years, RNN networks with LSTM blocks have been used extensively in machine learning tasks that process sequential data. These networks have proven to be particularly good at sequential language processing tasks by being more able to accurately predict the next most likely word in a given sequence than traditional neural networks. This study trained an RNN / LSTM neural network on three different translations of 150 biblical Psalms - in both English and Korean. The resulting model is then fed an input word and a length number from which it automatically generates a new Psalm of the desired length based on the patterns it recognized while training. The results of training the network on both English text and Korean text are compared and discussed.

키워드

Text Analysis, RNN, LSTM, Text Generation

Ⅰ. Introduction

In recent years, RNN networks with LSTM blocks have been used extensively in machine learning tasks that process sequential data. RNN networks cycle data in such a way as to allow previous nodes to affect the input of later nodes. And LSTM blocks solve the vanishing gradient problem often experienced when training RNNs alone by selectively remembering or forgetting certain pieces of data, thus prioritizing certain inputs into subsequent RNN nodes.

* corresponding author

Text prediction and text generation are sequential data modeling tasks that have become increasingly common and useful in recent years. Sequential data modeling tasks must be able to process inputs and outputs of varying lengths, as well as prioritize certain pieces of data by relying on "memories" processd by the LSTM blocks. With a well trained model, such neural networks are able to accurately predict and generate whole sequences of new text.

II. Related Research

Text generation has previously been performed for automatic image captioning[1] by using a CNN to classify or identify objects within an image, and then using an RNN / LSTM model to caption it. Text generation has also been demonstrated to generate a new script for a TV show[2], or even to write new Shakespeare-like text[3-4].

This research focuses on a comparison between English and Korean text generation using the same RNN / LSTM architecture, and the same data source (the 150 biblical Psalms), albeit with varying translations[5-10].

III. Dataset Collection & Cleaning

In order to increase the size and robustness of the datasets, three translations of the 150 biblical Psalms were gathered in both English and Korean. In English, these translations included the KJV (King James Version)[5], NIV (New International Version)[6], and ESV (English Standard Version)[7]. In Korean, these translations included the KLB (Korean Living Bible, 현대인의 성경)[7], KRV (Korean Revised Version, 개역한글)[8], and RNKSV (Revised New Korean Standard Version, 새번역)[9]. The translations were copied into text files from Bible.com, and line breaks were removed. Figure 1 displays a wordcount for each translation.

Version	Word count	Char count
KJV	44,442	235,496
NIV	41,414	219,793
ESV	44,085	229,173
KLB	28,833	122,394
KRV	26,256	109,094
RNKSV	32,337	139,899

Figure 1. Wordcount for each Bible translation.

All three translations for each language were then appended together into a single file containing 450 Psalms each. In English, the file contained 129,941 words. In Korean the file contained 87,426 words.

The datasets were then converted to lowercase, numbers were removed, punctuation was tokenized, and the individual words were split into a dictionary to determine a unique word count for each language. In English, there were 12,253 unique words with an average wordcount of 288.76 per Psalm. In Korean, there were 22,738 unique words and an average of 194.28 words per Psalm.

IV. Results & Analysis

Both languages used the same RNN/LSTM architecture, and initially, both models were trained with the same parameters. The input sequence length was 15 because that is the average length of an English sentence. The batch size was 256, learning rate was 0.001, and the number of epochs was 20. The network used 2 LSTM layers, an initial embedding layer with a dimension of 512, and a final linear layer with a dimension of 1024, as well as a dropout layer set to 0.45.

After training for 20 epochs, the models for each language were saved, and a generator was created by running the RNN on a given word. The topk() function was used to select the top 5 most likely next words in the sequence, and one of those was randomly selected to become the next input into the RNN and continue generating the Psalm.

Initially, the English model was saved with a loss of 1.31, and the generator produced relatively convincing new text. Figure 2 shows a sample.

```
['lord']
lord. do not let my heart be drawn to
what is evil so that i dwell in the
tents of kedar! too long have i lived
among those who hate peace.
```

Figure 2. An RNN generated English Psalm.

However, although the Korean model was initially saved with a similar loss of 1.56, the model took much longer to train and produced less convincing results. Figure 3 shows a sample.

['입에']
입에 해칠세라 사랑하심을 따라다니고 사랑하심을 따라다니고 따라다니고 포위하고 부모가 숲속에 포도나무가,

Figure 3. An RNN generated Korean Psalm.

The previous figure shows one word repeated multiple times. Therefore, the input sequence length was changed to 20, which is closer to the average length of a Korean sentence, and the number of epochs was increased to 60, and the model was trained again. This time, it produced much better results. Figure 4 shows a sample.

['여호와여']
여호와여 나를 보호하소서. 내가 이 재난이 지날 때까지 주의 날개 그늘 아래 피하겠습니다. 내가 가장 높으신 하나님께 부르짖노라. 그는 나를 위해 자기 뜻을 이루시는 분이시다.

Figure 4. Better RNN generated Korean Psalm.

Ⅴ. Conclusion & Future Work

While both models may not be perfect, it is noteworthy that Korean took significantly longer to train (almost double the time at 20 epochs), and initially produced far worse results. This probably has to do with the fact that the Korean text had significantly more unique words than English (almost double). Additionally, it is likely that the reason for the increased number of unique words in Korean is due to the spacing and use of particle markers such as 이/가, 은/는, 을/를, as well as other grammatical structures like conjunctions and possessives (-고, -지만, -의) that are not spaced away from the main nouns and verbs, nor separated by punctuation, such as the English apostrophe.

Therefore, a better Korean model might be trained if it could accurately decipher different parts of speech, or if parts of speech were additionally assigned to each word, perhaps by using a separate dictionary lookup function. Or, a better Korean model might be trained if the particles and grammatical structures could be tokenized separately from the words themselves as punctuation has been in English.

References

[1] Tan, Elaina, and Lakshay Sharma. “Neural Image Captioning.” arXiv, July 2, 2019. https://doi.org/10.48550/arXiv.1907.02065.

[2] Mangal, Sanidhya, Poorva Joshi, and Rahul Modak. “LSTM vs. GRU vs. Bidirectional RNN for Script Generation.” arXiv, August 12, 2019. https://doi.org/10.48550/arXiv.1908.04332.

[3] Ming, Yao, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. “Understanding Hidden Memories of Recurrent Neural Networks,” October 30, 2017.

[4] “Shakespeare Text Generation (using RNN LSTM).” Google Colaboratory. Accessed September 30, 2022. https://colab.research.google.com/github/trekhleb/machine-learning-experiments/blob/master/experiments/text_generation_shakespeare_rnn/text_generation_shakespeare_rnn.ipynb.

[5] KJV Bible | King James Version | YouVersion. Accessed September 30, 2022. https://www.bible.com/versions/1-kjv-king-james-version.

[6] NIV Bible | New International Version | YouVersion. Accessed September 30, 2022. https://www.bible.com/versions/111-niv-new-international-version.

[7] ESV Bible | English Standard Version 2016 | YouVersion. Accessed September 30, 2022. https://www.bible.com/versions/59-esv-english-standard-version-2016.

[8] KLB Bible | 현대인의 성경 | YouVersion. Sept. 30, 2022. https://www.bible.com/versions/86-klb-.

[9] KRV Bible | 개역한글 | YouVersion. Accessed Sep. 30, 2022. https://www.bible.com/versions/88-krv.

[10] RNKSV Bible | 새번역 | YouVersion. Accessed Sep. 30, 2022. https://www.bible.com/versions/142-rnksv.