ScciencePG
Science Publishing Group

Assisting Access to COVID-19 Information Through Deep Learning Based Machine Translation: Attention Mechanism Via Bidirectional GRU

Daniel Chang

North London Collegiate School Jeju, Jeju, Korea

Email address:
danielchangsh@gmail.com

To cite this article:

Daniel Chang. Assisting Access to COVID-19 Information Through Deep Learning Based Machine Translation: Attention Mechanism Via Bidirectional GRU. *American Journal of Data Mining and Knowledge Discovery*. Vol. 6, No. 1, 2021, pp. 9-15.
doi: 10.11648/j.ajdmkd.20210601.12

Abstract: Due to the recent COVID-19 crisis, there is an increasing need for effective communication and sharing of information internationally in various fields. One of the obstacles that these needs face are language: In texts such as COVID-19 related research, currently existing machine translations which are effective in normal texts because they are trained with normal-context data are often inaccurate, and manual translation is slow and laboursome. So, the exchange of information is being delayed. To overcome this language barrier, this project aimed to create a model that is effective for translating COVID-19 crisis related data specifically. In the research, there are two models created: one is trained with TAUS English-French Corona Crisis Corpus, and another used transfer learning by Kaggle English-French corpus and then trained with TAUS corpus. The model consisted of four bidirectional GRU layers, and used rmsprop as optimizer. The project evaluated the model using the BLEU score. The first model had a higher BLEU score than the second model, supporting the hypothesis that loosely related datasets decrease the quality of translation. In further research, evaluation on this model on different language pairs and use datasets in other specific fields will be conducted.

Keywords: Seq2seq, COVID-19, Bi-GRU, Machine Translation, NLP

1. Introduction

1.1. Background

Recently, COVID-19 became a great cause of concern around the globe, having an effect on many aspects of our society. COVID-19 is a disease caused by a newly discovered virus SARS-CoV-2, causing symptoms such as fever, cough, and fatigue. Due to its highly contagious nature, the disease rapidly spread across the world since its first reported in Wuhan on 31 December 2019, infecting over 198 million people and resulting in more than 4.2 million deaths. Though vaccination restrained the spread of disease, recent mutation of the virus increased the need for further development on the ways to deal with the pandemic [1].

Many people around the world are working on different fields to overcome the pandemic: medical staff are finding effective ways to treat patients, governments are employing policies to stop the spread of disease, and researchers are developing effective vaccinations or cures for the disease. It is essential for people around the world to share their developments on the ways to deal with COVID-19. However, these information is not shared quickly among the countries, partially due to language barriers (Coronavirus Disease [2].

Natural language processing is a branch of computer science that deals with the computers' ability to understand text or spoken words. Using artificial intelligence and deep learning, this technology is used in various fields including translation and is projected to grow in market size [3]. Below graph shows revenue from the NLP market worldwide from.
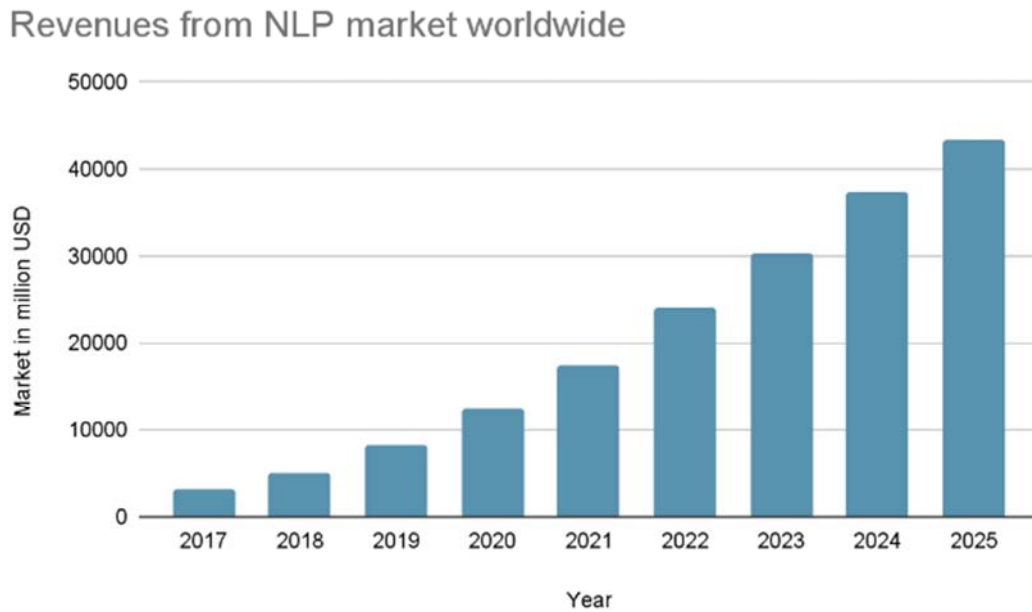
Revenues from NLP market worldwide

Figure 1. Revenues from NLP market worldwide from 2017 to 2025.

1.2. Objective

Translators currently in use, such as Google translator, are developed by training them with broad sets of data that are not specified in one subject. This allows high translation accuracy in normal-context texts, though it may result in inaccurate translation in specialized context texts like COVID-19 related articles. Such texts cannot be translated through former translators, so manual translation is done, which is very inefficient and laboursome. This project propose a new translation model trained specifically on data related to the disease that has higher accuracy than former translators and translation models in other works. The experiment consists of two parts. First one is constructing attention based translation models and training with the target dataset, which consists of COVID-19 related sentences. In the second experiment, this project apply transfer learning through another dataset which contains corpus for English - French translation. Then, evaluation on the both models through BLEU scores and comparison will be conducted.

2. Related Works

The research proposed a Neural Machine Translation (NMT) model specialized for translating COVID-19 related text in French, Spanish, German, Italian, Russian, and Chinese to English and English to these languages. The model was trained on TAUS Corona Crisis Corpora data set and evaluated on BLEU score. It resulted in higher BLEU score in every language compared to Google translator; English to Chinese (10.87) and Chinese to English (6.79) showed the highest score gap [4].

The research compared performances of Moses, a traditional statistical machine translation (SMT), and proposed recurrent neural networks (RNN) translation

(NMT). The models were trained by English-Hindi parallel corpus from MTIL2017. SMT resulted in better translation for long sentences while NMT was better in shorter sentences. NMT's BLEU score was 3.57 [5].

The researchers created a machine translation system trained explicitly on COVID-19 data. The translator is freely available to anyone to translate information about the disease in German, French, Italian, and Spanish into English and in opposite directions. The researchers showed that their MT engines outperformed other online MT engines. They also found that training every available dataset actually deteriorates the performance, as tested with five corpora: TAUS Corona Crisis Corpus, EMEA Corpus, Sketch Engine Corpus, ParaCrawl Corpus, and Wikipedia Corpus. Highest BLEU scores for the final MT engine were 61.49 in German-to-English and 54.41 in English-to-German [6].

The researchers proposed a theoretical framework using natural language processing to support medical staff. Through voice recognition, the system receives words spoken on it. Then, it processes the text and fetches information in the database related to it. This can aid medical staff in the pandemic by allowing them to provide information to the patients without physically contacting them [7].

The researchers created and trained Machine Translation systems that can translate from English to German, Greek, French, Italian, Spanish and Swedish, making COVID-related information quickly accessible to these language speakers. The team used three models: standard NMT training with back-translation, transfer learning, and multilingual training. The result shows that there does not exist a single model that functions best. For example, in English to German, BLEU score of transfer learning is 0.2 higher than that of standard training. On the other hand, in English to French, standard training's BLEU score is 0.6 higher than multilingual or standard training's score [8].

3. Materials and Methods

3.1. Data Description

Datas used in the experiment are from Taus Corona Crisis Corpora. They are generated from a collective industry charity effort in which participants contributed to expand the volume of the data and the number of languages. Matching Data selection to Datacloud and ParaCrawl data have also been used to generate the corpora. The corpora contains language pairs of English to French, German, Italian, Spanish, Chinese, and Russian [9].

Figure 2. COVID-19 related English - French corpus dataset from kaggle.

3.2. Data Description for Transfer Learning

The dataset for transfer learning is from Kaggle, which is available at: https://www.kaggle.com/devicharith/language-translation-englishfrench. It consists of 176000 English-French sentence pairs (Language Translation [10].

Figure 3. English - French corpus dataset from kaggle.

3.3. GRU

The recurrent neural network (RNN) has a slightly different architecture compared to the artificial neural network (ANN) or deep neural network (DNN). While the DNN never reuses the weight from the previous layer during the training, RNN exploits the weights from the previous layer for the input. Long short term memory (LSTM) and gated recurrent unit (GRU) are the representative models of the RNN [11]. LSTM is the advanced model from the vanilla RNN and consists of 3 gates which are forget gate, input gate, and the output gate. On the other hand, GRU comprises two gates, which are reset gate and update gate [12]. The update gate is a combination of the forget gate and the input gate from the LSTM and it determines the amount of the information to remember from the past and the present. The main purpose of the reset gate is to reset the past information. It utilizes the sigmoid function as an activation function in order to multiply the value (0,1) from the previous hidden layer [13].
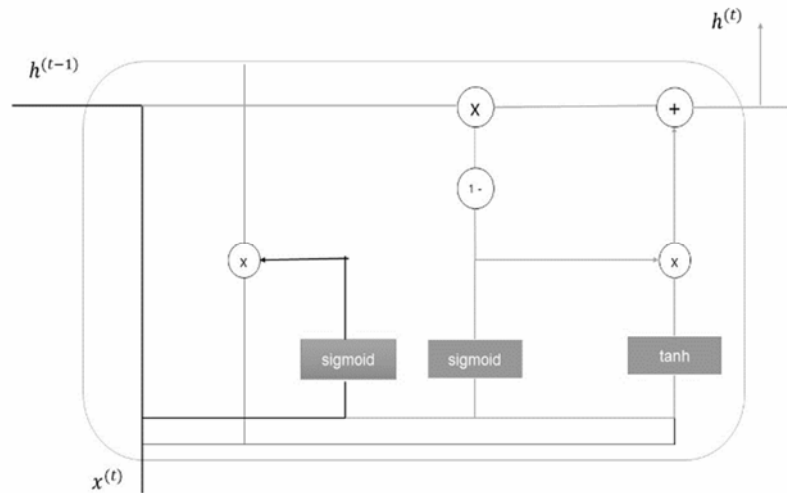
Figure 4. Overall architecture of GRU.

3.4. Sequence to Sequence

Sequence to sequence (seq2seq) model includes the encoder, context vector and the decoder. The role of the encoder is to encode the input data and the decoder is to decode the encoded data. When the encoder gets the input sentence, it extracts the cell state and the hidden state from the last sequences from the sentence. The context vector involves the information about the output from the encoder. Then, the decoder sequentially generates the output from the context vector. Both the encoder and the deoder mainly consist of the RNN networks and for better performance, LSTM and GRU are mostly applied [14].
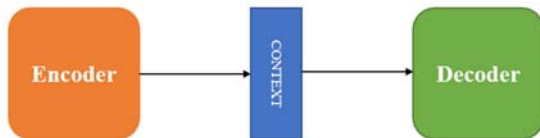

Figure 5. Overview architecture of seq2seq model.

3.5. Attention Mechanism

Attention mechanism was developed in order to solve the downside of the seq2seq model. As the information that the encoder passes to the decoder is a fixed vector, it can not contain all the information that is needed. The information that the encoder passes to the decoder is a fixed length vector. Fixed length means that no matter how long the sentence is, it always converts to a vector of the same length. Therefore, the attention mechanism improves the seq2seq model both in encoder and decoder parts. Unlike the encoder of the seq2seq model, the encoder in the attention mechanism utilizes the hidden state of all LSTM layers, not just the hidden state of the last LSTM layer in the encoder. This method allows us to figure out the constraint of the fixed length vector. Furthermore, for the advance of the decoder, it also receives an additional total of hidden state vectors. Through this, all the information about the sentences entered by the encoder was delivered to the decoder [15].
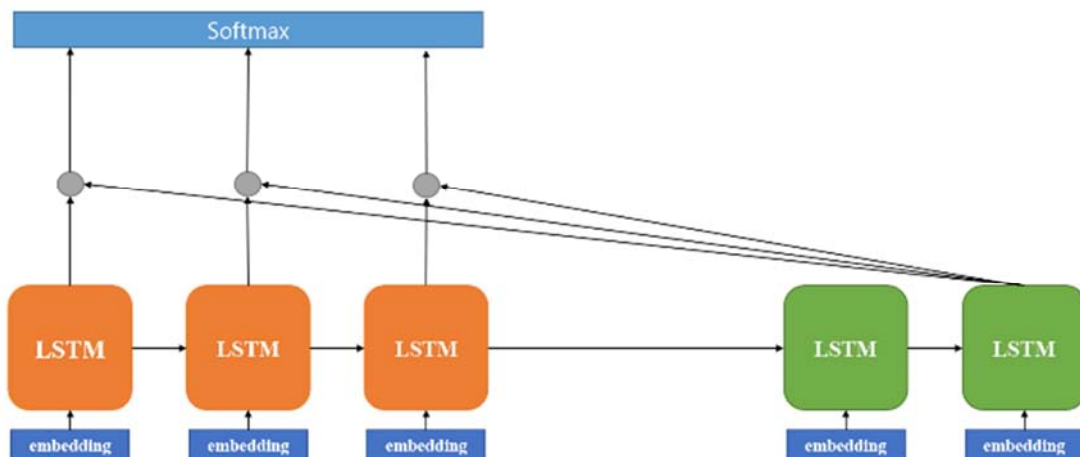

Figure 6. Overall architecture of attention mechanism.

3.6. Hardware Specs

1. CPU Model Name: Intel (R) Xeon (R)
2. CPU Frequency: 2.30GHz
3. Number of CPU cores: 2
4. Available RAM: 12GB
5. GPU Model Name: Nvidia K80
6. GPU Memory: 12GB
7. GPU Memory Clock: 0.82GHz
8. Performance: 4.1 TFLOPS
9. Available RAM: 12GB
10. Disk space: 358GB

4. Results

4.1. Result from the Fist Experiment

In the research, we modified the number and types of layers to optimize the accuracy of the translation. The bar graph shows the BLEU score of six models we tested. All six models used the optimizer as rmsprop and their hidden size was 256 nodes. Layers before the repeat_vector layer were for encoding, and ones after it were for decoding.

The model with the highest BLEU score used four bidirectional GRU layers. The BLEU score was 0.392. The next high-scoring model used six bidirectional GRU layers, with a BLEU score of 0.279. The third used two bidirectional GRU and two GRU layers, and had a score of 0.279. The fourth used four bidirectional LSTM layers, with a score of 0.253. The fifth model used two bidirectional LSTM and two LSTM layers, and had a score of 0.226. The one with the lowest score of 0.217 used two LSTM layers.

The line graph shows the change in loss during the training. The loss of the train set decreased nearly linearly from 2.35 to 1.31, while the loss of the validation set fluctuated while decreasing from 1.20 to 1.06.
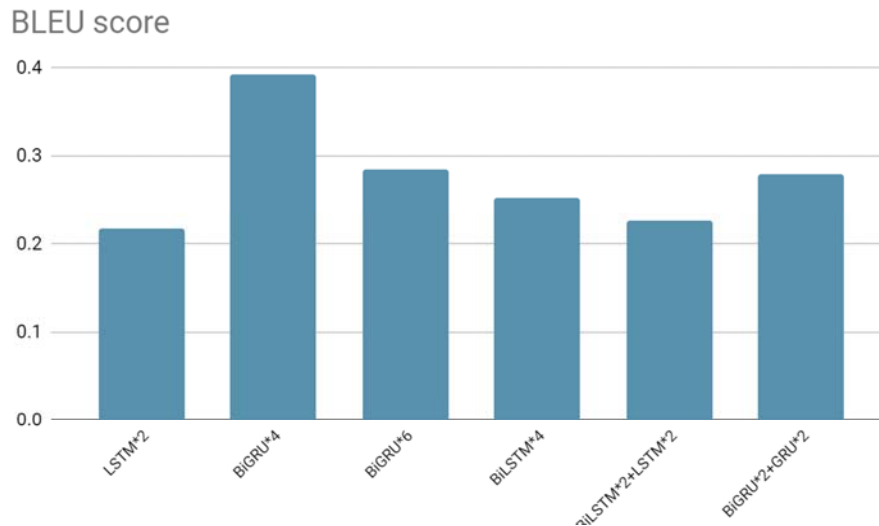

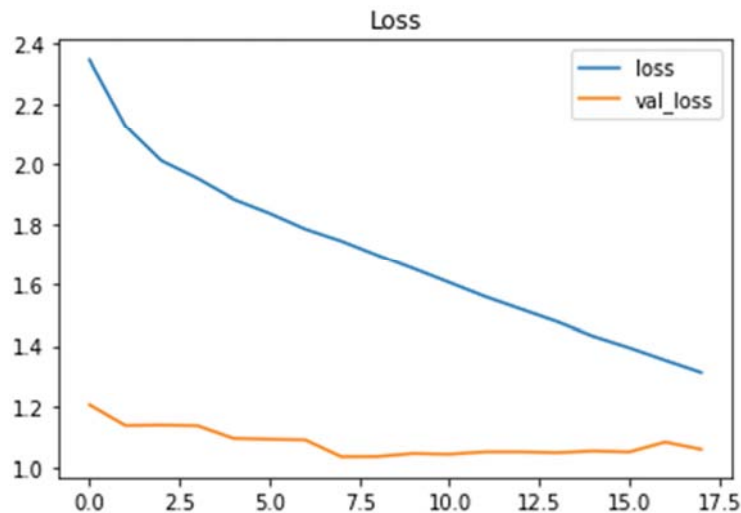**Figure 7.** *Comparison of BLEU score of various deep learning methods.*

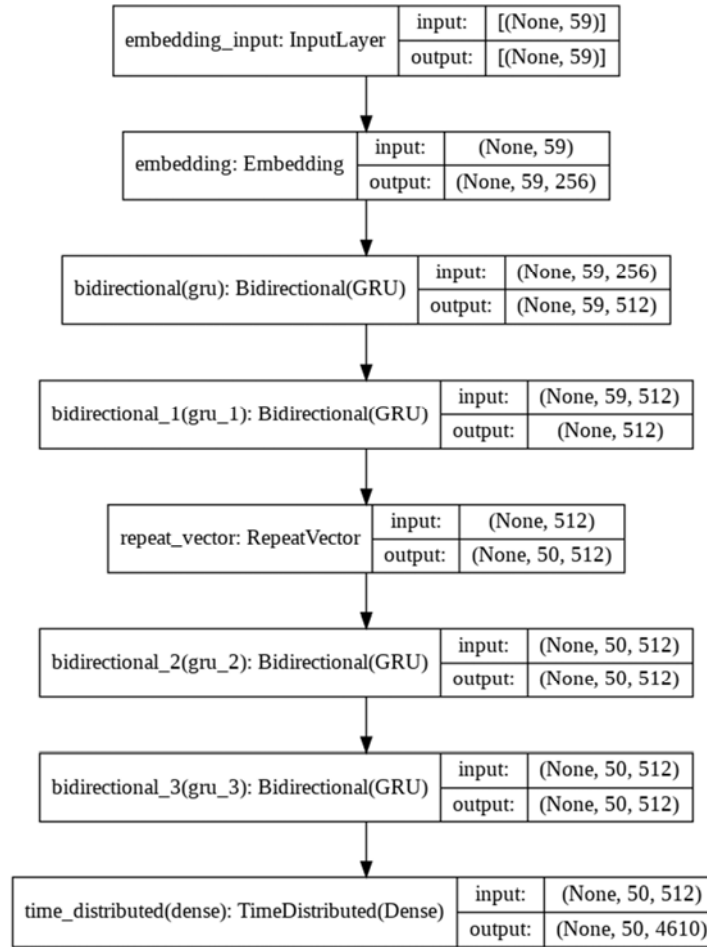**Figure 8.** *Loss and validation loss graph from the proposed model.*

***Figure 9.** Overall structure of our proposed translator model.*

4.2. Result from the Second Experiment

From the second experiment, the models which were also utilized in the previous experiment were pre trained with the other dataset. The model with highest BLEU score of 0.43 also used four bidirectional GRU; other models had BLEU score below 0.2. Overall, the models yielded a lower score than translation without transfer learning.
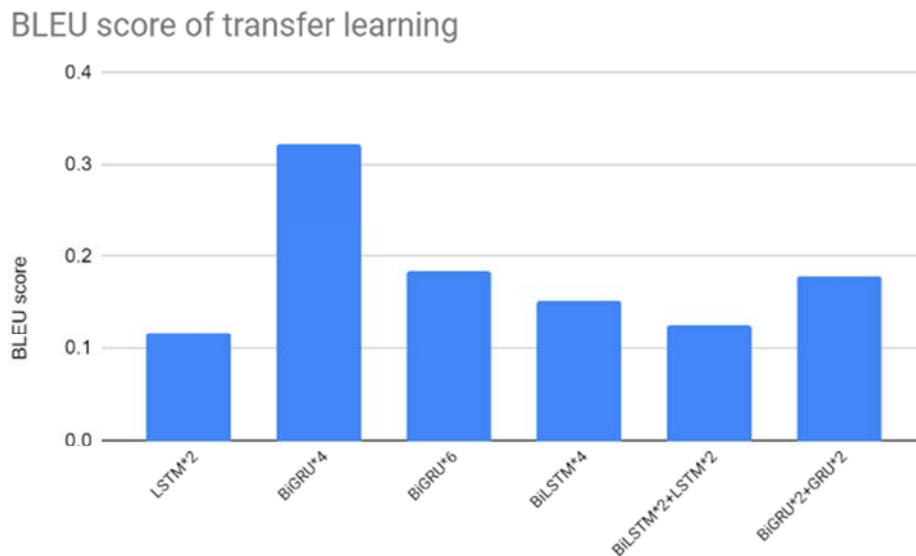

***Figure 10.** Comparison of BLEU score of various pre-trained deep learning models.*

5. Discussion

5.1. Principal Finding

The research resulted in relatively high performance in terms of BLEU score, despite the fact that the RAM available was only 12GB; the highest BLEU score was better than the score in other researches such as Mahata et al. When we trained a translation model on French-English data from kaggle, and trained it again using COVID-10 related data, the BLEU score was low. Transfer learning, usually used when the provided dataset is small, was not effective on data because it had distinct characteristics, such as professional medical terms. This shows how transfer learning is only effective when the data contains similar vocabularies and is in a similar context.

5.2. Limitation

The model was only trained and tested on English-French language pairs, though other related researches have done it on multiple languages. Because of insufficient hardware specs such as RAM size, the model had a limited amount of training, and we were unable to use the full dataset available. Therefore, in the further research, we will train the model we made on different datasets that similarly contained words used only on a specific profession, and test if it had higher BLEU score than other translators.

6. Conclusion

In this experiment, the model with four bidirectional GRU layers showed the highest BLEU score of 0.39 when trained and tested with English-French TAUS corona crisis corpus: this was higher than other researches in the same topic, despite the fact that small RAM size limited the number of training and the size of dataset available. When the model was trained through transfer learning, the score dropped to 0.32, showing that an unrelated dataset is not effective in training translators for specific purposes. The model was not tested in different language pairs other than English-French, so further research can be useful.

References

[1] Coronavirus (COVID-19). (2021, September 2). Google News. https://news.google.com/COVID19/map?hl=en-US&mid=%2Fm%2F06qd3&gl=US&ceid=US%3Aen.

[2] Coronavirus disease (COVID-19). (2020, October 12). World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-COVID-19.

[3] Revenue from nlp market worldwide - Google zoeken. (2020, June 8). Statistica. https://www.google.com/search?q=revenue+from+nlp+market+worldwide&oq=revenue+from+nlp+market+worldwide&aqs=chrome.69i57j33i160.4547j1j4&sourceid=chrome&ie=UTF-8.

[4] Park, C. J., Kim, K. H., Park, K. N., & Lim, H. S. (2020). Neural Machine translation specialized for Coronavirus Disease-19 (COVID-19). Journal of the Korea Convergence Society, 11 (9), 7-13. https://doi.org/10.15207/JKCS.2020.11.9.007.

[5] Mahata, S. K., Das, D., & Bandyopadhyay, S. (2019). MTIL2017: Machine Translation Using Recurrent Neural Network on Statistical Machine Translation. Journal of Intelligent Systems, 28 (3), 447–453. https://doi.org/10.1515/jisys-2018-0016.

[6] Way, A., Haque, R., Xie, G., Gaspari, F., Popović, M., & Poncelas, A. (2020). Rapid Development of Competitive Translation Engines for Access to Multilingual COVID-19 Information. Informatics, 7 (2), 19. https://doi.org/10.3390/informatics7020019.

[7] Komal, K., & Sharma, A. (2020). NATURAL LANGUAGE PROCESSING: AN APPROACH TO AID EMERGENCY SERVICES IN COVID-19 PANDEMIC. International Journal of Innovative Research in Computer Science & Technology, 8 (3). https://doi.org/10.21276/ijircst.2020.8.3.32.

[8] Kvapilikova, I., & Bojar, O. (2020). CUNI Machine Translation Systems for the COVID-19 MLIA Initiative.

[9] Corona Corpus - TAUS Matching Data. (n.d.). TAUS. Retrieved September 3, 2021, from https://md.taus.net/corona.

[10] Language Translation (English-French). (2020, April 8). Kaggle. https://www.kaggle.com/devicharith/language-translation-englishfrench.

[11] Cheon, M. J., Lee, D. H., Joo, H. S., & Lee, O. (2021). Deep learning based hybrid approach of detecting fraudulent transactions. Journal of Theoretical and Applied Information Technology, 99 (16), 4044-4054.

[12] Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. Journal of Artificial Intelligence and Soft Computing Research, 9 (4), 235–245. https://doi.org/10.2478/jaiscr-2019-0006.

[13] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078.

[14] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[15] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473.