Algerian Dialect Translation Applied on COVID-19 Social Media Comments

Amel Slim1, Ahlem Melouah1, Yousef Faghihi2, and Khouloud Sahib1

1 Department of Computer Science, Laboratory of Research in Computer Science (LRI), University of Badji Mokhtar, P.O. Box 12, 23000 Annaba, Algeria
Selimamel79@gmail.com, melouahlem@yahoo.fr, Sahibkhouloudil2@gmail.com
2 University of Quebec in Trois-Rivières, Trois-Rivières, QC, Canada
Usef.Faghihi@uqtr.ca

Abstract. This work is part of a study on the propagation of misinformation about COVID-19 and its impact on Algerian society. It studies the problem of Algerian dialect translation applied to COVID-19 social media communications. The proposed system begins by filtering messages to identify comments that talk about COVID-19. Then, COVID-19 texts are translated from the Algerian dialect to formal standard Arabic. The filtering process is based on the long short-term memory (LSTM) model. The translation process is based on the embedding-GRU model. Experimental results give precision rates of about 99.98% in the filtering process and about 97.56% in the translation process. The achieved BLUE score is 22.10.

Keywords: Translation · Algerian dialects · Covid-19 · Deep learning

1 Introduction

Covid-19 burst out in the city of Wuhan China and became pandemic very rapidly [1]. During this pandemic, people use social media to share opinions and perceptions about coronavirus. They exchange and acquire distinct types of information at a historic and unprecedented scale [2]. However, people’s opinion about coronavirus is the scariest scene because most information is untruth. According to the world health organization, untruth information is “spreading faster than the virus”. Misinformation is very dangerous on personality psychology; it increases the level of stress on themselves and with the others. It is very important to detect this type of information and to understand its impact on society. We must identify all messages who publish comments that cause terror in the hearts of others, as well as those who publish worn medicine COVID-19 treatments.

Arabic countries are in front of an important count of misinformation. Like all countries, they must control this propagation but first, they must surmount the dialect language barrier. Arabic countries’ peoples communicate using the language of the dialect as Levantine dialect, Egyptian, Algerian, Tunisian, etc. The Algerian dialect is a
difficult language to understand, even for those who are Arab but not Algerian. It is one of the most difficult dialects in Arabic countries.

Through this work, we are trying to provide clues on the problem of corona-virus information propagation in social media expressed with the Algerian dialect. Firstly, a filtering process identifies Coronavirus Algerian dialect messages. Secondly, a translation process transforms the informal message to a formal one. The filtering applies a long short-term memory model. The translation transforms the Algerian dialect sentences into Model Standard Arabic (MSA) forms. The main objective of this work is to make the Covid-19 Algerian dialect messages understandable.

The rest of the paper is organized as follows. Section 2 reviews related work done in the field of machine translation. Section 3 details the adopted filtering model, the proposed machine translation neural network and the used database. Section 4 presents the results with discussion. Finally, Sect. 5 concludes the paper.

2 Neural Machine Translation

With deep learning machine translation becomes more efficient. Many researches excelled in this field, and several deep learning models are proposed for machine translation.

The first neural machine translation model is an encoder-decoder model. The Encoder-Decoder is a recurrent neural network designed to solve sequence-to-sequence problems. This model gives more cleverness of any work in machine translation. Several articles have used the encoder-decoder model in the language translation problem. This section presents some of these articles.

[3] proposed an encoder-decoder model combined with a statistical machine translation approach; the model can either mark a pair of sequences or generate a target sequence from a source sequence. [4] translated the Algerian Arabic dialect into MSA (Modern Standard Arabic) through two steps; a transliteration step to convert the Arabic letter into the Arabizi letter, and a translation step to transform the Arabizi sentences into MSA. [5] demonstrated the performance of the encoder-decoder model by evaluating and comparing it to other models for an English-French translation problem.

Another area of research in the field of translation is transliteration. Transliteration consists in transforming a grapheme transcription from one writing system to another while preserving its pronunciation. [6] has shown that the encoder-decoder can solve a transliteration problem by using it for the Latin-Tunisian dialect transliteration.

The evolution of neural machine translation has not stopped at the encoder-decoder model; some works have added the attentional mechanism in neural machine translation. Attention, which is considered as an alignment model [7], offers the possibility to encode the relevant parts of the source sentence at each step of the translation. Several works have used this alignment principle in machine translation. The first attentional encoder-decoder model was proposed in [8]. The main idea of this model is to add the alignment mechanism to the basic concept of encoder-decoder for English-French translation. This idea has been developed further in two Arabic dialect translation systems [9]. The first system, dialectal translation to a standard language (D2SLT), is
based on the attentional sequence-to-sequence learning model. The second system, Google Neural Machine Translation (GNMT), is based on a sequence-to-sequence model with the addition of a residual connection, an attentional mechanism and a bidirectional encoder all proposed in [10]. [11] extended the basic encoder-decoder by including the alignment mechanism in the task of English-to-German translation.

3 Methodology

In social media, COVID-19 has spread terror among people because of rumors and myths about the virus. It is a problem for all societies, including Algerian society. Deep learning models have dominated the field of machine learning for natural language processing. Several works have used deep learning for classification and translation. In this work, we apply deep learning models to the translation of Covid-19 commentaries from Algerian dialects into a formal language.

Figure 1 describes the architecture of the proposed systems. Filtering and translation are the two main operations of the system. They are based on the deep learning model. In the last number of scholarly works, there have been remarkable developments in deep learning [12, 13]. Architectures such as a convolutional neural network (CNN), LSTM, and GRU have obtained competitive results in several competitions (e.g., computer vision, signal, and natural language processing) [14]. The Long Short-Term Memory networks “LSTMs”, introduced in [15], are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems and, they are widely used in different domains. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods is practically their default behavior.

The following sections are devoted to these operations, but first, we present all the used data.
A. Dataset

In this study, we construct a dataset using two sources. The first source is Corpora PADIC [16] dataset. The second source is Facebook.

Corpora PADIC contains five Arabic dialects and their translation to MSA (model standard Arabic). The Arabic dialects in the Corpora PADIC dataset are ANB Algerian, ALG Algerian, Palestine, Syrian and Tunisian. Corpora PADIC contains 6413 sentences of each dialect all collected before the COVID-19 pandemic. All ALG dialects sentences in PADIC (6413 sentences) compose the first part of our dataset. The second part of our dataset is composed of 1200 sentences collected from Facebook comments on the COVID-19 virus. Table 1 gives some examples of the used dataset.

Table 1. A sample of the dataset with their MSA and English translations.

| Algerian dialect sentence | MSA translation | English translation |
|--------------------------|-----------------|---------------------|
| ندريق يختبر بتلكورونا النكبة الآن سوف سوف يصيبون بتلكورونا النكبة الآن سوف | Now they will be infected with corona. | Now they will be infected with corona. |
| يصيب أحيانا وربما ماأنه لا يفهمون كيفهم | The problem is they will infect other people. They do not understand like animals. God suffices us. | The problem is they will infect other people. They do not understand like animals. God suffices us. |
| لم يعتقدوا أن الفيروس موجود في الأشياء التي لم يعتقدوا أن الفيروس موجود في الأشياء التي ليس لها سماح فيلام بفيروس راء فالنحو لفسوم | They did not think that the virus was attached to the things that they had touched | They did not think that the virus was attached to the things that they had touched |
| للعالم يلته بيفن موانا ليس فقط في البلدة. أيق في للعالم يلته بيفن موانا ليس فقط في البلدة. أيق في | The whole world is burying its dead not only in Blida. Stay in your home | The whole world is burying its dead not only in Blida. Stay in your home |

B. Filtering

Deep learning is the concept adopted for the filtering. LSTM model classifies sentences into two categories: sentences with COVID-19 information and sentences without COVID-19 information as shown in Fig. 2.
The filtering model architecture starts with an embedding layer to quantify and categorize the semantic similarities between linguistic items based on their distributional properties in large samples of linguistic data. Then, the dropout layer forces the LSTM to learn useful robust features in conjunction with different random subsets of other neurons. Later, the LSTM layer followed by another dropout layer classifies the input data. Finally, a batch normalization layer and a dense layer improve the performance, speed, and stability of our architecture. Table 2 shows the number of the parameters of each layer.

**Table 2.** Filtering model layers and parameters.

| Layer                | Output shape | Parameters |
|----------------------|--------------|------------|
| Embedding            | (None, 55, 32) | 455808     |
| Dropout              | (None, 55, 32) | 0          |
| LSTM                 | (None, 250)  | 283000     |
| Dropout              | (None, 250)  | 0          |
| Batch normalisation  | (None, 250)  | 1000       |
| Dense                | (None, 1)    | 251        |

C. Translation

A neural machine translation system is any neural network that maps a source sentence, \(X_1, \ldots, X_N\), to a target sentence, \(Y_1, \ldots, Y_T\), where all sentences are assumed to terminate with a special “end-of-sentence” token \(<\text{EOS}>\) [17].

In translation stage, sentences about COVID-19 are translated from the Arabic Algerian dialects into MSA. The translation applies an embedding gated recurrent unit (GRU). Typically, the network architecture content embedding input layer, GRU layer, dropout layer, two time-distributed layers, and dropout layer. The translation network accepts one dialect in input and gives its translation in output as shown in Fig. 3.

In a translation system, the words or phrases from the source vocabulary are mapped to vectors of real numbers. It yields word embedding as the weights of the first layer, which is usually referred to as the Embedding Layer. Conceptually it involves a mathematical embedding from space with many dimensions per word to a continuous vector space with a much lower dimension. In parallel, Gated Recurrent Neural Networks (Gated RNNs) have shown success in several applications [18]. They were proposed by [19] to make each recurrent unit to adaptively capture dependencies of different time scales. Similarly, to the LSTM unit, the GRU has gating units that modulate the flow of information inside the unit, however, without having separate memory cells [20].
We designed the parameters of the model as shown in Table 3. This table shows the number of parameters in each layer.

| Layer          | Output shape | Parameters |
|----------------|--------------|------------|
| Embedding      | (None, 52, 256) | 3130368    |
| GRU            | (None, 52, 256) | 393984     |
| Time-distributed | (None, 52, 1024) | 263168      |
| Dropout        | (None, 52, 1024) | 0          |
| Time-distributed | (None, 52, 10600) | 10865000   |

4 Result and Discussion

The translation of Covid-19 commentaries written in the Algerian dialect is the problem addressed by this study. The strategy is to filter the comments between those referring to Covid-19 and those not referring to Covid-19. Next, translate the comments referring to Covid-19 from the Algerian dialect into a formal language. An LSTM model ensures the filtering and a GRU model does the translation. We use Tensorflow and Keras to implement the proposed system. The model evaluation engages

- Filtering: accuracy metric, the binary cross-entropy loss, and the Adam optimizer. With 50 epochs and batch of size 1024 sequence for each iteration.
- Translation: accuracy metric, sparse categorical cross-entropy loss, and Adam optimizer. With a learning rate $\alpha = 0.005$, 30 epochs, and a batch of size 1024 sequence for each iteration.
The performances are graphically depicted with accuracy and loss. Figure 4 and Fig. 5 present the results of the experiments. Each figure depicts the accuracy and the loss of each architecture.

**Fig. 4.** Filtering model performance, at left the loss of the model and at right the model accuracy.

**Fig. 5.** Translation model performance, at left the loss of the model and at right the model accuracy.

An overall loss score and accuracy based on the train, validation, and test dataset are computed and used to determine the model’s performances. Table 4 presents the metrics results.

| Model   | Accuracy | Loss |
|---------|----------|------|
|         | Train    | Validation | Test | Train | Validation | Test |
| Filtering | 99.94    | 99.99     | 91.96 | 0.0010 | 3.3902e-04  | 0.48 |
| Translation | 97.25    | 97.56     | 0.14  | 0.13   |             |      |
We tested the proposed system on 14026 sentences. The training step used 13000 sentences and, 10% of these sentences were reserved for the evaluation. The rest, 1026 sentences, were exploited in the test step. We selected the weight averaging method to aggregate the individual class metrics. The selected hyperparameter values are shown in Table 4. In the translation step, 1300 sentences were used for the training step, and, 10% of these sentences were used in the evaluation step.

An interesting result is the values of the BLUE scores of Fig. 6.

Table 5 shows the results of the filtering process. The good performance of the proposed model is evident; all the sentences are well classified.

Table 5. Examples of the filtering process. (0: sentences not about COVID-19. 1: sentences about COVID-19.)

| Algerien dialect example | Target | Prediction |
|-------------------------|--------|------------|
| ...                     | 0      | 0          |
| كمية النازعين والاستهزاء عند البعض فالتوقعات الامكانيات على روايح حاول علي الناس في الالوان  | 1      | 1          |
| هذا فيروس معدى بدء عباد الله  |         |            |
| الي والدي مترفون موعش مغني بالجر الصحي  | 1      | 1          |

Table 6 gives the results of the translation of certain comments from the Algerian dialect into MSA. We observe that the proposed model translated wrongly the words "لناشأ" when it should have been translated by the words "لناشيء". However, the proposed model gives a correct answer to the sentence "هككش كورونا بلجزائر خرنطي هذه مسرحية"; this sentence is correctly translated.
To prove the performance of the proposed system, we compared it with the system proposed in [4]. Table 6 presents the results of this comparison. Although the proposed embedding GRU translation architecture is simpler, it is clear that it gave better results than the encoder-decoder model proposed in [4] (Table 7).

### Table 6. Examples of the translation process.

| Source Reference | Reference translation | MAS translation | English translation |
|------------------|-----------------------|-----------------|---------------------|
| There is no corona in Algeria. Everything is a Myth | There is no corona in Algeria. Everything is a Myth | There is no corona in Algeria. Everything is a Myth | It is a dramatization. |
| When you suspect someone is injured, please stay away from him as much as possible | You saw someone was injured, please stay away from him as possible | عادل الشهيد في إصابة أحد أفراد مركز بعد إلقاءه على قيد الامكان | It is a dramatization. |
| Many of the citrus fruits, such as lemon and orange, increase the immune system | Many of the citrus fruits, such as lemon and orange, increase the immune system | Many of the citrus fruits, such as lemon and orange, increase the immune system |

### Table 7. Translation comparison results.

| Translation model | Dataset | BLEU score |
|-------------------|---------|------------|
| Embedding GRU proposed model | PADIC + 12 00 sentences | 22.10 |
| Encoder-decoder RNN [4] | PADIC | 4.26 |

5 Conclusion

In this work, we were interested in the translation of comments written in the Algerian dialect on the subject of the Covid-19 pandemic. A two-step system is used to recognize sentences that mention Covid-19. Then, it translates them into a formal language. The recognition phase is based on a filtering principle realized by a deep learning model (LSTM). The phase of translation of the Algerian dialect into MSA is based on an embedding GRU model. The filtering model has reached an accuracy score of about 91.96%. Besides, the translation model obtained 22.10 BLUE points.

Acknowledgments. We are grateful to the Direction Générale de la Recherche Scientifique et du Développement Technologique (DGRSDT) which kindly supported this research, as well as to the Laboratoire de Recherche Informatique (LRI) where this study was conducted.
References

1. Wu, J.T., Leung, K., Leung, G.M.: Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modeling study. Lancet 395(10225), 689–697 (2020)

2. Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.L., Duan, W., Tsoi, K.K.F., Wang, F.Y.: Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on weibo. IEEE Trans. Comput. Soc. Syst. 7(2), 556–562 (2020)

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

4. Guellil, I., Azouaou, F., Abbas, M.: Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In: The 31st Pacific Asia Conference on Language, Information and Computation PACLIC, vol. 31, p. 2017 (2017)

5. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)

6. Younes, J., Souissi, E., Achour, H., Ferchichi, A.: A sequence-to-sequence based approach for the double transliteration of Tunisian dialect. Procedia Comput. Sci. 142, 238–245 (2018)

7. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)

8. Ghader, H., Monz, C.: What does attention in neural machine translation pay attention to? arXiv preprint arXiv:1710.03348 (2017)

9. Farhan, W., Talafha, B., Abuammar, A., Jaikat, R., Al-Ayyoub, M., Tarakji, A.B., Toma, A.: Unsupervised dialectal neural machine translation. Inf. Process. Manag. 57(3), 102181 (2020)

10. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J.: Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)

11. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)

12. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

13. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Inf. Fusion 42, 146–157 (2018)

14. Gonçalves, S., Cortez, P., Moro, S.: A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Comput. Appl. 32(11), 6793–6807 (2019)

15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)

16. Meftouh, K., Harrat, S., Smaili, K.: PADIC: extension and new experiments. In: 7th International Conference on Advanced Technologies (ICAT 2018), Antalya – Turkey April (2018)

17. Wu, L., Tian, F., Qin, T., Lai, J., Liu, T.Y.: A study of reinforcement learning for neural machine translation. arXiv preprint arXiv:1808.08866 (2018)

18. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th International Mid-West Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600. IEEE, August 2017
19. Gulcehre, C., Cho, K., Pascanu, R., Bengio, Y.: Learned-norm pooling for deep feedforward and recurrent neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 530–546. Springer, Heidelberg, September 2014

20. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)