



# A Proposed Sequence-to-Sequence Modeling for Arabic Dialect Machine Translation

## Dialect Arabic (DA) to English Translator

Sawsan Al-Odibat

Jordan University for science and technology

Irbid, Jordan

[smalodibat15@cit.just.edu.jo](mailto:smalodibat15@cit.just.edu.jo)

Yasmin Al-Sayyah

Jordan University for science and technology

Irbid, Jordan

[Yosayaheen14@cit.just.edu.jo](mailto:Yosayaheen14@cit.just.edu.jo)

### Abstract

Neural machine is considered as an innovative approach to translate and make statistical machine model for translation that rely only on neural networks. The basic translation models using neural machine often contain encoder and decoder operations. The encoder makes cuttings a fixed-length sentence representation from a variable-length sentence that are input for translation, but the decoder produces right translation for this fixed-length representation. This paper motivates to compare between recurrent neural traditional networks that are recursive based (RNN), enhanced once long short-term memory unit (LSTM) and sub type of LSTM that are gated recurrent unit (GRU). The process sequence-to-sequence labeling check recurrent unit types to make translating between language pairs, for our study from Arabic Levantine to English. The perplexity measure is used to score the effect the trained model. BLEU scores is used to measure the quality of the translation.

**Keywords**—*recursive neural networks, machine translation. sequence to sequence modeling, dialect Arabic*



## I. INTRODUCTION

Modern Standard Arabic (MSA) is not typically spoken, instead people communicate in their native Arab dialects, such as Levantine, that are spoken at the regional level of countries Jordan, Syria, Palestine, and Lebanon. These dialects have no standard orthography (insufficiently written) and written without short vowel diacritics, resultant in substantial ambiguity [15]. Thus, dialects do not have orthographic representation of short letters and need a high degree of words sense of disambiguation. Moreover, an Arabic written dialect is often mixed with MSA, numbers, Arabizi as well as other languages. Further, these dialects might overlap with each other like MSA and English words written in Arabic letters, however, they vary in their vocabulary, pronunciation, morphology, and word order. This leads to emerge Arabic Natural Language Processing (ANLP) [16].

The nature of the Arabic language is complex and requires dealing with many issues of structure. There are a few Arabic NLP Tools in the fields of machine translation, Arabic named entity recognition and sentiment analysis [15] [16]. Because the need for such tools is urgent, we are motivated to develop a new approach for machine translation to translate Arabic dialect to English directly or going through MSA. The most essential problem is when Arabic Dialect texts include many complicated co-references, translated and transliterated named entities whose spelling in general tends to be inconsistent with MSA.

We aim at developing ANLP system for the dialects by first extracting and classifying the systematic grammatical features of a dialect, making it more like MSA and then applying MSA natural language processing tools to process a text. Then the translation process from MSA to English can take place. The primary goal is to extract the features of the dialect and deploy them in the process of translation for easy conversion from type to type and format to format.

Particularly, the problem of machine translation between Arabic Dialect and English is proposed in this research. We propose a two-way language model: from the Levantine variety of DA to English. The pre-processing of the input text (dialect text) is processed through word embedding (vectorization) and vocabulary building. The proposed approach attempts to improve the alignment and generate improved translated output via sequence to sequence modeling.

## II. BACKGROUND

Before several years ago, machine translation has received a good attention as an active research topic. Formerly, machine translation approaches have been developed as dictionary-based systems that depend on the grammatical rules, words order and words location. Over years, the researchers have added a knowledge



www.mecsjs.com

model to improve their results. Many models have been developed to improve machine translation models, such as statistical models. Eventually, other methods have emerged and became leading into a machine learning field such as reordering, filtering, and alignment. In the last decade, rule-based approaches have been concentrated in machine translation systems for a long time. However, natural language processing currently is going towards neural networks [2].

Dislike statistical and probabilistic machine translation models, Neural Machine Translation (NMT) constructs a single join model to improve the performance of translation. Mostly, NMT models belong to those models encoding and decoding the source sentence or sequence into a fixed length vector that can generate the target translation [1]. Deep Neural Networks (DNNs) have received good attention due to their effective performance in achieving complex machine learning activities. In contrast, DNNs cannot map sequence-to-sequence modeling, although they work well in large datasets [11].

The Recursive Neural Network has proven its fruitful role in achieving machine translation tasks with high accuracy. Recently, RNNs have played a critical role in getting more attention and interest of researchers in machine learning, specifically machine translation [2].

#### A. Long Short-Term Model (LSTM)

Recurrent neural networks (RNN) are networks that contain loops inside them, letting information to keep on. Loops mark as a kind of mysterious, RNN look like a duplicate of the one network, each output of the loop is considered input to next loop [9]. RNN used for sequences and lists of data because of the nature of its architecture. The problems that RNN can solve are: translation of any languages, speech recognition, image captioning and others. Connect preceding information to the current task that is concept Long-Term Dependencies, which are not supported on traditional RNN when consideration of long gap between relevant information [12].

Long short term memory network (LSTMs) is a kind of RNN that do many tasks with better performance than RNN. LSTM use past information to handle a current task. LSTMs take in consideration long gap between relevant information in order to perform current task. [14] LSTMs work extremely well on a huge problem and now extensively used. The default behavior of LSTMs is Remembering information for extensive periods. Traditional RNNs have a very simple repeating module that has a very modest structure like a single layer and a single neural. LSTMs have such chain structure; however, repeating module in LSTMs has dissimilar structure to RNN structure. Instead of single neural layer, LSTMs have four interacting layers in very unusual way [14].

Each line has an entire vector that outputs from the previous node and its output goes to others node. The pink circles characterize as point wise to make operations. However, yellow boxes are considered as learned



www.mecsjs.com

layers. The Lines merging mean concatenation, even though the line forking means content copies and transfers copies to many locations [9]. The idea of LSTMs that it contains cell state represented as horizontal line appears running. The information flows fairly on Cell state along the way it unchanged [9].

### B. (GRU)

Gates are places where information through based on some conditions. In LSTMs. there are three gates: forget gate as shown on figure 2, where sigmoid layer controls information that pass throw cell state, the output of cell state ranges from zero to one; when the value becomes great, means that it will through the gate otherwise it will be forgotten [14]. The input sentence in a variable length is encoded into a fixed length of representation in NMT models, as well as this representation is decoded to generate the correct translation [3].

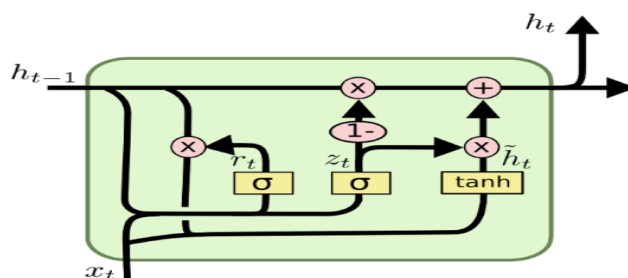


Figure 1: GRU unit gates[12]

Update gate interested in the information values that will update based sigmoid layer value and tanh layer the create an updated vector of new values that candidate as shown in figure 2. Next step, combine result from sigmoid layer and tanh layer to make new update state [14]. Output gate as shown on figure 2, decide information that is going to output. The thing that will output is based on current cell state, but it will also have filtered. First, we apply sigmoid layer that chooses parts of cell state that will output. Tanh layer apply to cell state so specific part only outputs [14]. Gated Recurrent Unit is a variant of LSTM, the idea comes from (Cho et al. in 2014a). The modification is combines between the cell state and the other hidden as show in figure 2. The gate stills the same but the flow data between them is changing [12].

## III. LITERATURE REVIEW

A good approach using neural networks integrated with a statistical machine translation called RNN Encoder-Decoder was proposed by Cho et al., in 2014. It composes two RNNs; one for encoding and the other is for decoding. The approach can encode each sequence of symbols in arbitrary length and represent it into a fixed-length vector, and, reversely, it converts the representation into the corresponding targeted symbols. A novel hidden unit was invented to control and update the results during reading and generating any sequence. To increase the conditional probability of training, the proposed model joins the both encoder



[www.mecsjs.com](http://www.mecsjs.com)

and decoder, as well as the conditional probabilities combined with RNN empirically improves the performance of a statistical machine translation system on word and phrase pairs basis. More specifically, the approach showed an improved learning of semantic and syntactic analysis of linguistic phrases. BLUE scores have been improved by the RNN Encoder-Decoder that also improves Statistical Machine Translation (SMT) performance with the neural net language model. However, the architecture of the proposed approach did not replace any part of phrase table to let proposing the target phrases, and it has a further open research to be applied to other applications, like speech transcripts [2].

In a 2014 paper by Bahdanau et al., a proposed extended model to the fixed length vector of the basic Encoder-Decoder called RNN-search by ignoring the form of the source segment and enabling an automatic soft-search of the source sequence to predict the target word, which can improve the performance of machine translation. One of the apparent features of the proposed model is the capacity of treating sentences in any length even longer sentences. The innovative approach achieved an improved comparable performance in translating English to French carried out by phrase-based approaches. It also achieves English-to-French translation by extending the traditional Encoder-Decoder architecture to search of word inputs to generate the target word, overcoming the problem of the approach of generating fixed-length vector. Further, the model focuses on the relevant information close to the next generated target word, which reveals to have a positive effect on the results of NMT with respect for long sentences. To test RNN-search model, an experiment has been conducted to prove its ability to translate from English to Arabic regardless the length of given sentence. In other words, the proposed model can correctly generate translation by aligning each target word to its corresponding word or annotation. However, the proposed approach cannot treat rare words such that a better understanding of the unknown words is needed to develop [1].

A new RNN Encoder-Decoder model called end-to-end NMT was proposed by Luong et al. to resolve the problem of inability of previous approaches to translate rare words. To address this problem, NMT system was trained on data (aligned words) to allow the system to determine the position of the word in the target sentence mapped to the associated word in the source sentence for each Out-Of-Vocabulary (OOV) word. In post-processing step, the information of this mapping can be used to translate each OOV word using identity translation or dictionary. The proposed approach utilized the strengths of phrase-based system to be adaptive for translating rare words. The technique of NMT system can achieve the state-of-the-art performance and not only to the deep LSTM for machine translation. The experiment was conducted on a dataset for translation from English to French that revealed to have an essential enhancement in terms of BLUE measure up to 37.5 points [11].

More recently, a paper by Jean et al. was proposed for machine translation to handle the problem of a large vocabulary and complexity with the aid of neural networks. The approach is based on sampling to break up the training complexity and decoding complexity that increase with the increasing number of target



www.mecsjs.com

words. Through selecting a subset of a large vocabulary of target words, the method can improve performance with no effect to the training complexity. The model was trained to show its performance, and it empirically achieved a comparable performance outperforming the state-of-the-art models like LSTM when it ensembles a very large vocabulary into a few models. The experiment was applied on English to German translation and English to French translation. The main feature of this approach is the reduction of the computational complexity in normalization the probability of the output words based on neural language models. The results showed an enhanced BLUE measure value up to 1 point (21.59) comparable to other approaches (that achieved 20.67) [7].

Grid Long Short-Term Memory (GLSTM), an enhanced network of LSTM cells, was introduced by Kalchbrenner, et al. to apply multidimensional grid on the sequences or vectors in a high dimensional data. The architecture of the network of LSTM cells compromises a set of connected cells along the layers that represent data dimensions. The model was applied to two empirical tasks: sequence memorization and algorithmic tasks. The results proved the effectiveness of the performance of Grid LSTM that uses 2D or more (N-way) translation model (Re-encoder) and applied to Chinese to English translation. Grid LSTM allows cells to communicate to other layers (multi-way interaction) compared to the regular connection of the base-line approaches. A flexible and powerful approach proposed in the paper has been applied to machine translation, character prediction and image classification. The proposed approach outperformed the phrase-based approach and showed strong advancement of performance [8].

Lately, a new attention-based method to improve NTM was focused on exploring the architecture of the source sentence. Two effective and simple classes: global and local in the attention-based mechanisms were examined in the paper. Global approach always looks at the positions of the words in the source sentence, while local approach looks only once a time at a portion of these words. Both approaches were examined to determine their effectiveness in machine translation from English to German bidirectionally. Extensive experiments were conducted to assess the proposed model with respect to handling long sentences, alignment quality, learning and the resultant translation. Local attention has achieved a significantly improved gain of BLUE measure up to 5 points against non-intentional systems. Indeed, the model yielded a novel result with 25.9 BLUE points as an improvement to the existing systems. Further, the proposed attention-based model has been compared to other non-intentional approaches, and it was superior in handling long sentences and translating names [10].

Multilingual NMT was proposed in [5] to enable one Neural Network Model (NTM) to pass some parameters that can translate from one language to another in a linearly growing. A single attention method is shared among several pair of languages, and a multi-way training (multilingual model) was implemented with multiple encoders and decoders. A single neural network mechanism can handle multi languages of source and target. The experiment was tested on a large scale of parallel corpora containing five languages. A





[www.mecsjs.com](http://www.mecsjs.com)

clear improved performance was observed over different models that were applied on each pair of languages (between low-resource languages). In conclusion, an improved translation quality of the language pairs was observed significantly. However, a large vocabulary tricks can be ensembled and applied by this multilingual model for further improvements.

An end-to-end sequence modeling was presented by Sutskever et al. based on multi-layered LSTM that encodes the input sequence into a fixed dimensional vector, in turn another LSTM decodes the output sequence from this vector. The approach was applied on English to French translation with 34.8 BLUE score, compared to phrase-based SMT system with 33.3 BLUE score. LSTM has no difficulty to handle long sentences and it further achieved 36.5 BLUE score. Phrases and sentences that are sensitive to word order can be learned by LSTM, therapy the reverse order of the words of the source sentence can improve its performance during translation and transition from source and target sentence with regard of short-term dependencies among them. More improvement to the approach are available for research even though it achieved success in sequence learning problems. However, due to limited memory, LSTM could report poor performance on long sentences. It is straightforward, simple model but it is a challenging task in sequence to sequence problems [13].

The success of RNN model and its recent encouraging results in many fields of natural language processing, motivated Greenstein and Daniel to propose an RNN model for machine translation from Japanese to English. The proposed model performs well to a given small vocabulary and small parallel corpus to adapt complexities of grammar. The training of translation model on large corpus is not possible within a limited time since these experiments need intensive time. A subset of the corpora was selected to train RNNsearch on variant sentence vocabulary and structure to obtain good model. Different datasets were modeled by the trained model that can extrapolate the sample sentences with similar vocabulary and structure to get high accuracy with the exact translation [6].

The properties of NMT, including standard RNN Encoder-Decoder model and gated hidden RNN, were discussed in the paper of Cho et al., 2014 NMT performs well on unknown words and short sentences, but it does not perform well when the number of unknown words and the length of the sentence are increasing. The analysis of two evaluated models showed that the performance of translation is affected by the vocabulary size. Moreover, two models can generate the correct translation mostly. Similarly, it can learn without supervision of the source language syntactic structure. However, the training of a neural network can be scaled up in both memory and computation to support larger vocabularies and the length of sentences (source and target). The model was established to mimic input sentence in terms of grammatical structure with no supervision of language syntactic structure that make the model apt for natural language processing more than machine translation [3].



#### IV. MEHTODOLOGY

The contribution of this paper falls into two sections: first is to use neural networks in MT applications, and the second is to evaluate the results of RNN using GRU in translation from DA to English. Using RNN Encoder-Decoder, the words are converted into 2-D embedding of words based on embedding matrix. In RNN Encoder-Decoder, a continuous representation of words is generated. In addition, the scores of RNN Encoder-Decoder approach can improve the overall performance of translation according to BLUE measure.

The trained model can realize the linguistic regularities at several levels of language: phrase level and word level (sequences). More natural language related applications can benefit from this approach in improvement and analysis of machine translation approaches.

##### A. Experiment settings

The proposed approach in this paper is interested in estimation of the performance of GRU sequence modelling in Machine Translation. The source code was obtained from the GitHub website that is developed based on TensorFlow. We made some changes to the code such as reducing some default parameters. All experiments were done on Ubuntu 16.04 machine with 8 RAM and core i7 processor.

##### B. Data set

The evaluation focuses on sequence to sequence modeling on a dialect Arabic dataset taken from [1]. It contains 1200 tweets in Syrian dialect. We divided it into three segments: training set (800 tweets), development set (200 tweets), and testing set (200 tweets), the data provided by Dr.Mohamad al-Smadi in Jordan university of science and technology.

Because our dataset (1200 tweets) is not big enough to implement enhanced sequence-to-sequence modeling with TensorFlow source code. the following changes help us to be more flexible in dealing with training time and development. For this reason, the number of units has been reduced to 512 rather than 1024, the batch size was reduced to 32 instead of 64, and the number of layers was cut down to 2 rather than 3. Further, the vocabulary size was reduced from 40000 to 30000. By default, GRU is implemented by the source code and we can change it to LSTM if the argument is changed.

#### V. RESULTS AND DISCUSSION

The evaluation of the proposed model was based on the task of dialect Arabic to English translation. The performance of the NMT is assessed by Blue test which was 44.7% as shown in below table

<b>Perplexity</b>	<b>GRU</b>
-------------------	------------





	<i>Step 800</i>	<i>Step 1000</i>	<i>Step 1200</i>
73.4	0.09564 2	0.10903 1	0.2144 56
25.9	0.16039 7	0.29634 5	0. 32724 1
13.06	0.18111 5	0.34536 7	0.4475 23

## VI. CONCLUSION

This paper proposed a novel approach in machine translation and evaluates the performance of sequence to sequence modeling to translate dialect Arabic to English. We make in evaluating models based on encoder–decoder performance and sentence-to-sentence labeling. The performance analyses of BLEU scores that work more efficiently when apply GRU type.

This paper motivates to make comparison between recurrent neural traditional networks that are recursive based (RNN), enhanced once long short-term memory unit (LSTM) and sub type of LSTM that gated recurrent unit (GRU). The process sequence-to-sequence labeling, check recurrent unit types to make translating between language pairs, for our study from Arabic Levantine to English. The perplexity measure used to score effect the trained model was 13.06%, while BLEU scores used to measure quality translation was 44.75%.

## REFERENCES

- [1] B, Dzmitry, K, Cho, and Y, Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473*. 2014.
- [2] C, Kyunghyun, Bm Merriënboer, C, Gulcehre, D, Bahdanau, F, Bougares, H, Schwenk, and Y, Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078*. 2014.
- [3] C, Kyunghyun, B, Merriënboer, D, Bahdanau, and Y, Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259*. 2014.
- [4] Ch, Junyoung, C, Gulcehre, K, Cho, and Y, Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555*. 2014.



www.mecsjs.com

- [5] F, Orhan, K, Cho, and Y, Bengio. "Multi-way, multilingual neural machine translation with a shared attention mechanism." *arXiv preprint arXiv:1601.01073*. 2016.
- [6] G, Eric, and D, Penner. "Japanese-to-English Machine Translation Using Recurrent Neural Networks." 2015.
- [7] J, Sébastien, O, Firat, K, Cho, R, Memisevic, and Y, Bengio. "Montreal neural machine translation systems for WMT'15." In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 134-140. 2015.
- [8] K, Nal, I, Danihelka, and A, Graves. "Grid long short-term memory." *arXiv preprint arXiv:1507.01526*. 2015.
- [9] K, Andrej, J, Johnson, and L, Fei-Fei. "Visualizing and understanding recurrent networks." *arXiv preprint arXiv:1506.02078*. 2015.
- [10] L, Minh-Thang, H, Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025*. 2015.
- [11] L, Minh-Thang, I, Sutskever, Quoc V. L, Oriol Vinyals, and W, Zaremba. "Addressing the rare word problem in neural machine translation." *arXiv preprint arXiv:1410.8206*. 2014.
- [12] O, Christopher. "Understanding LSTM Networks." URL <http://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-chain.png>. 2015.
- [13] 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.
- [14] Y, Kaisheng, et al. "Spoken language understanding using long short-term memory neural networks." *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014.
- [15] I, Abraham, and S, Roukos. "A maximum entropy word aligner for arabic-english machine translation." In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 89-96. Association for Computational Linguistics, 2005.
- [16] H, Nizar, N, Zalmout, D, Taji, H, Hoang, and M, Alzate. "A Parallel Corpus for Evaluating Machine Translation between Arabic and European Languages." *EACL 2017* 2017: 235.