DIA-English-Arabic neural machine translation domain: sulfur industry

Diadeen Ali Hameed[1], Tahseen Ameen Faisal[2], Alaa Khudhair Abbas[3], Harith Abdullah Ali[4], Ghanim Thiab Hasan[1]
[1]Department of Electrical Engineering, Shirqat Engineering College, Tikrit University, Tikrit, Iraq
[2]Department of English Language, College of Basic Education, Tikrit University, Tikrit, Iraq
[3]Department of Petroleum Control System, Oil Processing Engineering College, University of Tikrit, Tikrit, Iraq
[4]Department of Civil Engineering, Engineering College, Tikrit University, Tikrit, Iraq

Article Info

Article history:

Received Mar 18, 2022
Revised Jun 17, 2022
Accepted Jul 4, 2022

Keywords:

DIA translator
English-Arabic translator
Google translator
Machine translation
Natural language processing
Neural approach

ABSTRACT

The aim of this paper is the design and development a new English-Arabic neural machine translation (NMT) called DIA translation system. The main purpose of the designing system is to study translator limited sulfur industry domain as a stand-alone tool in order to improve the translation quality. Machine translation (MT) are very sensitive to the domains they were trained on and can be integrated with general (English-Arabic) MT systems. The proposed system has mainly four directions: supports chemical symbols, terms, phrase, and text and it is evaluated by using (1,200) various English declarative sentences which written by English language experts. The obtained results indicate that this system is high effective and has an accuracy of 79.33% in comparison with Google translator which has 38.67% for the same test samples.

Corresponding Author:

Ghanim Thiab Hasan
Department of Electrical Engineering, Faculty of Shirqat Engineering, Tikrit University
Salah-Alden Tikrit, Iraq
Email: ganimdiab@yahoo.com

1. INTRODUCTION

English is a universal language that is widely used in the science [1] as well as in the technology fields. English-to-Arabic neural machine translation (NMT) is particularly important and is mainly based on the transfer classification. Comparatively little work has been done on machine translation (MT) systems involving Arabic language as the source or target language [2]. MT method based on neural networks has several advantages over the other approaches, which is one of the most widely explored areas in MT system [3]. The system of MT are considered as very sensitive to domains that they are trained on because each domain has its specific style, terminology and sentence structure [4]. Ambiguity of words is often a problem in machine translation systems [5]. For example, the English word “frequency” must be translated differently if it occurs in a technical or economic context. The main idea of our work is based on the fact that the neural models can benefit from domain information to select the most appropriate sentence terms and structures, while using information from all areas to improve basic translation quality.

Hadla *et al.* [6] reported that MT technology in the field of neural network throughput in machine translation systems is an important area of research to optimize [7] the efficiency and modesty of the sulfur industry through side barriers. We've extended this idea to domain management. Our goal is to enable models with different training data to produce translations within the domain [8]. This means extending general

NMT models to specific areas and their specific concepts and styles [9] without compromising the quality of translation of more general information.

Previous work has shown that the NMT model can investigate attention distributions that intuitively explain the reasonable correlation between source and target languages [10]. Literature in this area indicates that a little works have been conducted in the Arabic language as a target language. Statistical machine translation (SMT) has been the main translation paradigm for decades [11]. Even before the advent of direct machine translation of neurons, neural networks were successfully used as a component of SMT systems. Perhaps one of the most significant experiments involved the use of a common language model to study sentence presentation [12], which led to a dramatic improvements in sentence-based translation and extended sentence systems.

Many new techniques have been proposed to improve MT for example manage the results of rare words, various attention mechanisms [13] and minimize sentence loss. Some tecent works also have especially dealt with domain adaptation for NMT by providing meta-information to the neural network. The present work is in line with this kind of approach, and translation accuracy of this system is gratifying, recent work has focused on adapting NMT domains in particular by providing metadata to neural networks [14]. So, the topic of this paper is a part of this approach. The power of the neural network in issues related to the decoder; the topics are varied and consist of product categories labeled with people [15]. Include thematic modeling of encoder and decoder components. The number of standard items is automatically extracted from the linear discriminant analysis (LDA) training model; each word in the sentence gets its own thematic vector. In our work, we also provide metadata and information about the domain [16].

2. METHOD

NMT is a technique based on neural networks and conditional probability of a sentence translated from the source language into the target sentences [17] which is widely used in the area of deep learning for MT. Sequence-by-sequence architecture is used in machine translation models to find the relationship between two different language pairs [18]. The system architecture (NMT) is shown in Figure 1. The algorithm used for performing English to Arabic translation can be explained with the help of the diagram shown [19]. Figure 2 illustrates the architecture of of the proposed DIA translator system.

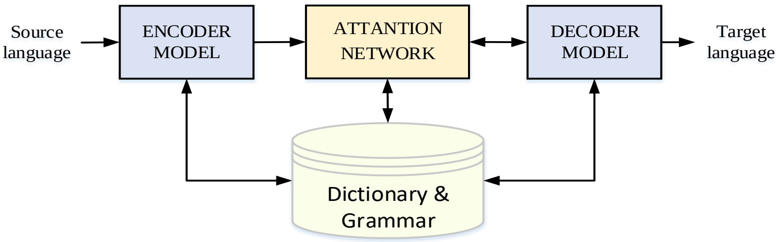

Figure 1. Architecture of the neural translation machine system

2.1. Encoder model

The source language analysis (English language) and data entry methods are appropriately prepared for machine translation. The input is a raw material for the whole system; a text file contains a well-structured collection of sentences written in English language. The effectiveness of source language analysis can be increased in the three steps applied to the morphological analysis, syntax analysis (parser tree) and semantic analysis [20].

The original words are first drawn with word vectors and then inserted into a double neural network (RNN) that reads the input letter S = {w1, w2, w3, …, Wn}; receives one element of the input string at each step, processes, collects, and disseminates information about that element. The coding part contains information that connects string chains with vector spaces to perform neural network calculations. Since words also have a meaningful sequence, a repetitive neural network is suitable for this task, the problem with this method is that it does not completely solve grammatical complexity, especially when translating the word nth into ocular language, RNN considered only (1..n)-word in the original sentence, but the grammatical meaning of the word also depends on the order of the words before and after the sentence: Using a two-dimensional model allows us to enter the meaning of past and future words to create an exact vector for the encoder output: but then it becomes a challenge, which word should we focus on? [21]. prepared a document showing that we can learn words in the language of the eye to focus by storing the previous result in long short-term memory (LSTM) units, then sorting according to each appropriate and selecting words with the highest scores.

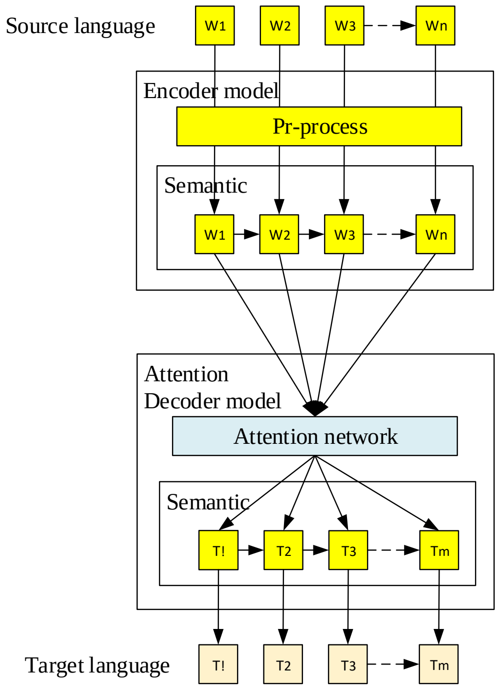

Figure 2. Architecture of the DIA translator system

2.2. Attention decoder model

Encoder-decoder models with attention have been proposed and then become a de-facto standard in the neural machine translation [22]. This part explains target language generation (Arabic language) and how output texts are appropriately translate on machine translation system. The effectiveness of generate equivalent target language can be increased in the two steps applied to the sentence reorder and semantic [23]. Assume B as a target sentence, in the decoding process, the following word is assumed using predefined words and units of (1) Target objects A = {A1, A2, A3,…., Time}. Using a chain rule, the distribution of expressions can be subtracted from left to right, since the focusing system is part of the neural network. Determine the components of the eye that are most important for each step of the decoder. At this point, the encoder does not need to squeeze the entire eye into a vector, it provides an indication of all flashing signals.

$$P(y = \mathrm{y}|x = \mathrm{x}) = \prod_{t=1}^{T} p(yt|y0, \dots, yt - 1, x1, \dots, xs) \qquad (1)$$

NMT models which conform the (1) is referred to as L2R autoregressive NMT [24]-[26], for the prediction at time-step t is taken as a input at time-step t+1. The model uses the attention of a series of coding, and the weights determine the attention of relationships that combine information from different places. This framework is very appropriate for our current study because we emphasize the ability of NMT to collect contextual dependencies from a broader context beyond sentence boundaries.

Focus is chosen to target a subset of the hidden encoder states per target word. The model first generates a p (t) alignment position for each target word at time t, while learning the alignment positions in attention. In other words, it enables efficient GPU-based training and decoding with a mini series and determining whether the translation order is different from the original sentence (original word 1 can be words 4in a translated sentence).

The following algorithm includes three main steps used in the machine translation (MT) system:

1st step: encoder network.
- Input (source text).
- Semantic source text.

2nd step: attention-decoder network.
- Target text generation.
- Optimize the target text.

3rd step: evaluation and rank.
- Evaluation DIA translator with Google translator.
- Rank DIA translator and Google translator from best to worst evaluation.

Each step for the encoder or decoder is inputs and generating output for that time step. In this resides the limitation of classic sequence to sequence models; the encoder is “forced” to send only a single vector, regardless of the length of input source and the model over fits with all sequences.

3. RESULTS AND DISCUSSION

This study has been conducted on bases of the data set constructed by (sulfur production company catalogs), the (1,200 sentences) of the previous dataset, that are divided by each reference translation of the sentences of all English-Arabic in the dataset of (4) main sentence functions, (text, terms, phrase, general text) with each of all English-Arabic sentence reference translations in the data set. The results of average precision for each phrase in the corpus of DIA translator and Google translator and are illustrated in Table 1 and Figure 3.

Table 1. Human evaluation average precision for each type

MT/Criterial	Terms by domain	Phrase by domain	Text by domain	Text without domain	Average precision
DIA MT system	0.85	0.80	0.73	0.61	0.793
Google translator	0.33	0.44	0.39	0.75	0.387

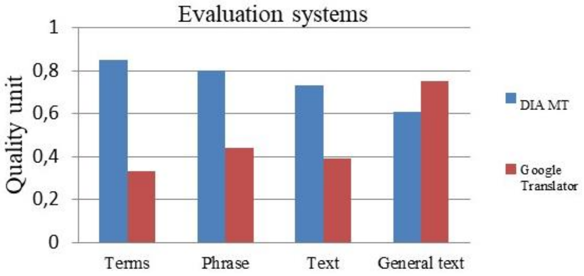

Figure 3. Average precision of evaluation systems

Research evaluation is most vital in considering success and failure of research work done so, this study uses sulfur industry domain by evaluation its system by specialized English-Arabic translation center at university of Tikrit vs. Google translator. The comparison between DIA translator with Google translator indicate that:

- Google translator doesn't support chemical symbols, while the (DIA translator) system supports chemical symbols in detail.
- In case of the (terms), it was stated that DIA translator scheme is much better than the Google translator. The reason for this is that the DIA translator system database contains translation file of English-Arabic terms.
- In case of the (phrase by sulfur industry domain), it can be seen that the Google translate system is capable of displaying the MT DIA navigation system in most cases.
- In the sulfur web industry, it can be seen that the Google transfer system is in most cases inferior to the MT DIA navigation system; Results testing from Google's translation system shows that the most sought-after analysis of these items is genuine and irreparable.
- It should also be noted that while Google translator cannot translate complex sentences with sulfur energy with accuracy, the 100 DIA translator interpreter can interpret some of these sentences on the phone.
- In general texts (texts that do not use sulfur energy), we note that the precision is the same in some simple sentences. Google's translation system is a much wider application than the Arabic machine translation system for multiple articles. Composite sentence structure.

Finally, it can be seen that the Google translate system is capable of displaying the MT DIA navigation system in most applications, as illustrated in Table 1. From the obtained results, we can conclude that the DIA translator is produces optimized outputs better than Google translator for sulfur industry domain (0.387 for Google and 0.793 for DIA translator) in the tests conducted.

4. CONCLUSION

In this work, the domain sulfur industry into a NMT for one of the most difficult language pairs (English-Arabic) has been used. From the results above obtained, it can conclude that the domain with byte pair encoding and pre-trained word embedding can performs better translation than the English-Arabic languages general translation techniques. The results obtained also indicate that the DIA MT system accuracy is approximately 79.3% compared with the submission accuracy for the Google translator which is approximately 38.67% in case of using domain sulfur industry. Finally, from all the results obtained, we can conclude that the DIA translator is fairly good accuracy and able to outperform many baseline translation systems.

FUTURE WORK

Since domain classification is a document level task, it would be interesting to extend the current study to document level translation.

REFERENCES

[1] T. A. Mesallam *et al.*, "Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/8783751.

[2] S. R. Shareef and Y. F. Irhayim, "A review: isolated arabic words recognition using artificial intelligent techniques," *Journal of Physics: Conference Series*, vol. 1897, no. 1, p. 12026, May 2021, doi: 10.1088/1742-6596/1897/1/012026.

[3] Y. K. Hussein, D. A. Hameed, L. I. Kalaf, B. Rahmatullah, and A. T. Al-Taani, "Automatic evaluating Russian-Arabic machine translation quality using BLEU method," *Revista AUS 25,* pp. 155-162, 2019.

[4] L. Schillingmann, J. Ernst, V. Keite, B. Wrede, A. S. Meyer, and E. Belke, "AlignTool: The automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes," *Behavior Research Methods*, vol. 50, no. 2, pp. 466–489, Jan. 2018, doi: 10.3758/s13428-017-1002-7.

[5] D. A. Hameed, Y. K. Hussein, L. I. Kalaf, and B. Rahmatullah, "A review on automatic machine translation approaches," *Revista AUS 26,* pp. 95-103, 2019.

[6] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative study between METEOR and BLEU methods of MT: Arabic into English translation as a case study," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 6, no. 11, 2015, doi: 10.14569/IJACSA.2015.061128.

[7] J. McKechnie, B. Ahmed, R. G. Osuna, P. Monroe, P. McCabe, and K. J. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review," *International Journal of Speech-Language Pathology*, vol. 20, no. 6, pp. 583–598, Jul. 2018, doi: 10.1080/17549507.2018.1477991.

[8] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of voice pathology detection and classification on different frequency regions using correlation functions," *Journal of Voice*, vol. 31, no. 1, pp. 3–15, Jan. 2017, doi: 10.1016/j.jvoice.2016.01.014.

[9] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491-79509, 2020, doi: 10.1109/ACCESS.2020.2990434.

[10] A. Kourd and K. Kourd, "Arabic isolated word speaker dependent recognition system," *British Journal of Mathematics & Computer Science*, vol. 14, no. 1, pp. 1–15, Jan. 2016, doi: 10.9734/bjmcs/2016/23034.

[11] A. Vaswani *et al.,* "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 371-383, 2016, doi: /10.1162/tacl_a_00105.

[13] L. Boussaid and M. Hassine, "Arabic isolated word recognition system using hybrid feature extraction techniques and neural network," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 29–37, Nov. 2018, doi: 10.1007/s10772-017-9480-7.

[14] R. S. Khudeyer, M. Alabbas, and M. Radif, "Multi-font arabic isolated character recognition using combining machine learning classifiers," *Journal of Southwest Jiaotong University*, vol. 55, no. 1, 2020, doi: 10.35741/issn.0258-2724.55.1.12.

[15] P. Kumar, H. Gauba, P. Pratim Roy, and D. P. Dogra, "A multimodal framework for sensor-based sign language recognition," *Neurocomputing*, vol. 259, pp. 21-38, Oct. 2017, doi: 10.1016/j.neucom.2016.08.132.

[16] A. Elsayed and H. Hamdy, "Arabic sign language (Arsl) recognition system using HMM," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 11, 2011, doi: 10.14569/IJACSA.2011.021108.

[17] Y. Faisal and A. M. Khalaf, "Speech recognition of isolated Arabic words via using wavelet transformation and fuzzy neural network," *Computer Engineering and Intelligent Systems*, vol. 7, no. 3, pp. 21–31, 2016, Accessed: Jan. 20, 2022. [Online]. Available: https://iiste.org/Journals/index.php/CEIS/article/view/29405.

[18] E. K. Elsayed and D. R. Fathy, "Semantic deep learning to translate dynamic sign language," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 1, pp. 316-325, 2021, doi: 10.22266/ijies 2021.0228.30.

[19] D. A. Hameed, T. A. Faisal, A. M. Alshaykha, G. T. Hasan, and H. A. Ali, "Automatic evaluating of Russian-Arabic machine translation quality using METEOR method," *AIP Conference Proceedings,* 2022, doi: 10.1063/5.0067018.

[20] S. Abdel, A. A, Rabouh, F. A. Elmisery, A. M. Brisha, and A. H. Khalil, "Arabic sign Language recognition using kinect sensor," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 15, no. 2, pp. 57-67, 2018, doi: 10.19026/rjaset.15.5292.

[21] M. ElBadawy, A. S. Elons, H. A. Shedeed, and M. F. Tolba, "Arabic sign language recognition with 3d convolutional neural networks," *In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE,* 2017, pp. 6671, doi: 10.1109/INTELCIS.2017.8260028.

[22] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," *In Proceedings of the SPECOM*, 2005, vol. 1, no. 3, , pp. 191–194.

[23] S. R. Shareef and Y. F. Al-Irhayim 'Towards developing impairments arabic speech dataset using deep learning,"*Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, pp. 1400-1405, 2022, doi: 10.11591/ijeecs.v25.i3.pp1400-1405.

[24] S. Abdel, A. A.-Rabouh, F. A. Elmisery, A. M. Brisha, and A. H. Khalil, "Arabic sign language recognition using kinect sensor," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 15, no. 2, pp. 57-67, 2018, doi: 10.19026/rjaset.15.5292.

[25] M. H. Ismail, S. A. Dawwd, and F. H. Ali "Dynamic hand gesture recognition of Arabic sign language by using deep convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 25, no. 2, February 2022, pp. 952-962, doi: 10.11591/ijeecs.v25.i2.pp952-962.

[26] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, "An automatic arabic sign language recognition system (Arslrs)," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 470-477, 2018, doi: 10. 1016/j.jksuci.2017.09.007.

BIOGRAPHIES OF AUTHORS

Diadeen Ali Hameed is a lecturer in the electrical engineering department, Tikrit university, Tikrit, Iraq. He received the B.Sc. from I am a Assistant Professor of Electrical Engineering at University of Tikrit, Iraq. He is an associate professor at the department of electrical engineering, Al-Sherqat engineering college, Tikrit University, Iraq, where he has been a faculty member since 2006. He graduated with a first-class honours B.Eng. degree at University of Mosul/Iraq and the M.Sc. degrees from University of Yurmouk, Jordin, all in Computer science and information technology. His research interests are in the area of computer and electrical engineering. He can be contacted at email: diaa@tu.idu.iq.

Tahseen Ameen Faisal is an Associate Professor at the Department of English language, basic educational college, Tikrit University, Iraq, where he has been a faculty member since 2000. He graduated with a first-class honours B.Eng. degree in, in 1988, and M.Sc. in Electrical Engineering fromBelgrade University, Serbia in 2000. His research interests are primarily in the area of Linguistics and Translation. He can be contacted at email: Tahseen.faisal@tu.edu.iq.

Alaa Khudhair Abbas is an Associate Professor at the College of oil processing engineering college, Department of Petroleum Control System Eng., University of Tikrit, Iraq, where he has been a faculty member since 2010. He graduated with a first-class honours M.Sc degree in in 2015. His research interests are primarily in the area of computer science Artificial inelegance. He can be contacted at email: alaa.programmer12@tu.edu.iq.

Harith Abdullah Ali is an Associate Professor at the Department of civil Engineering, Engineering College, Tikrit University, Iraq, where he has been a faculty member since 2006. He graduated with a first-class honours B.Eng. degree in electrical and Electronic Engineering from Tikrit Universityy, Tikrit, in 1998, and M.Sc. in 2005. And Phd. In 2017. His research interests are primarily in the area of civil engineering. He can be contacted at email: dr.harith@tu.edu.iq.

Ghanim Thiab Hasan is an Associate Professor at the Department of Electrical Engineering, Al-Sherqat engineering college, Tikrit University, Iraq, where he has been a faculty member since 2006. He graduated with a first-class honours B.Eng. degree in electrical and Electronic Engineering from Belgrade Universityy, Serbia, in 1984, and M.Sc. in Electrical Engineering from Belgrade University, Serbia in 1986. His research interests are primarily in the area of electrical and electronic engineering. He can be contacted at email: ganimdiab@yahoo.com.