Study and evaluation of “Seq-2-Seq” model competency in AI-based educational chatbot for the Marathi language

Mohammed Waseem Ashfaque*
Department of Computer Science & I.T. Dr. B.A. M. University ,Aurangabad, India
Charansing N. Kayte
Department of Digital and Cyber Forensics, Govt. Institute of Forensic Science, Aurangabad, India
Sohail Iqbal Malik
Department of IT Buraimi University College, Buraimi, Oman
Shaikh Abdul Hannan
Department of Computer Science and Information Technology, Albaha University, Albaha, Saudi Arabia.

Abstract:-

Nowadays, chatbot has become more popular in almost all fields, specifically after the Corona pandemic. In the education field, it has a potential impact, it becomes an assistant not only for students but for the teacher also. And is not only a blessing for students,and tutors, but also for educational institutes, and educational administrators. The implementation of the chatbot can be either rule-based or generative-based but using AI generative-based chatbot, can be more intelligent and smart, and friendly in many ways, that’s why it become more popular among educational entities.Which,“Seq-2-Seq” (“Seq-2-Seq”) model using recurrent neural network (RNN), is quite centric for developers and researchers in implementing artificial intelligence-based chatbots, though still it is in the early stage of its progress and development. Actually hosted for machine translation-based devices,This Sequence to Sequence model has been amended for conversational modeling containing questioning and answering machines like chatbots.

Conversely,the study and exploration of ideal sites of the different modules of the “Seq-to-seq” model for expected answering generation are tricky and partial. Moreover, there have been no more trials conducted to acknowledge how the “Seq-2-Seq” model deals with various questioning sets posted to it to create the expected answers. Our investigation intensifies toward the realistic assessmentofthe “Seq-2-Seq” modelwork and provides certain perceptions of these questions. Moreover, we understandhow an actual dataset can be designed and developed and how the questions are intended to practice and evaluate the competencies of a “Seq-2-Seq”-based question-answering model.

Keywords:- Chatbot, Recurrent neural networks, “Seq-2-Seq” model, AI, Marathi Language.

1. Introduction:-

The ability for chatbots to act as web-based trainers that can support students and respond to their inquiries and questions is an intriguing idea with potential intelligent machines that can speak with people in natural language and provide answers as question-answering chatbots. Chatbots that possess knowledge and reasoning seem to be capable of scale considerably more quickly than human employees. This viewpoint suggests that using chatbots in educational domain might be viewed as a resource for a learner-centered strategy. The manner that teaching and learning are done is changing as a result of the increased usage of modern technology. Although chatbots in education are not a recent development, little research has been done in this area. Since 2006, numerous studies have focused on using of chatbots to facilitate “teaching-learning” from many angles [1, 7]. Similar to one-on-one interactions between students and teachers, chatbots offer an interactive learning environment. Chatbots can serve education goals in more ways than only answering questions or distributing information among students. One such benefit is that they can help with the issue of individualized customer care for customers like students, parents, and alumni. The growing popularity of chatbots is also a result of their promise to save costs byAn important tool in a personalized learning environment designed to increase student involvement and collaboration is the chatbot or AI conversational tool. It makes it possible for students to assimilate information at their own pace. Level, not just in traditional classroom settings but also through distance learning [4], [7]. The development of artificial intelligence has made it possible for educators to offer each student a personalized learning environment. Additionally, chatbot technology has been shown to be a useful aid for first-year students in reducing their information load and fostering a sense of social connection with their professors [5].Chatbots can help educators in several ways, in their opinion. It is used as a tool for “masscommunication” to transmit “text messages” like notifications and reminders [5]. To make the most of this feature, students' smartphones should include a chatbot. A chatbot can also assist with duties linked to “homework” and “assignments”, like finding “spelling and grammar” errors, reviewing “homework”, offering “group projects”, and monitoring each student's progress and accomplishments [7].Byexamining the transcripts of their chatbot through discussions, teachers can assess how their pupils are progressing [10].

Established templates and rules created by tools like Amazon Lex and Google's Dialog Flow andmachine learning techniques like neural networks, chatbots may be produced based on previous samples of the common communication network that requires less effort. A chatbot is considered more intelligent if it can provide answers to unknown questions using methods like similarity testing, answer deduction, and response generation. The “Seq-2-Seq” model, which is built on an encoder-decoder framework powered by “Recurrent Neural Networks (RNN)”, is a commonly studied “model” for chatbot development [11]. As with other neural network models, “Seq-2-Seq” presents a No. of settings as well ashyper parameters must be modified to produce a functioning system that performs well. Examples of settings and hyper parameters that must be specified.What ties philosophy and artificial intelligence together? Even though the model was trained on the first set of questions, it should be capable ofanswer both of them properly.

Table 1. Research questions

Q.No	Questions
RQ1	Which encoding forms best uses keywords and characters for chatbots in education?
RQ2	How does the performance of the education chatbot change when failure rates are applied?

RQ3	What kinds of education chatbots can the Sequence to Sequence model, which is builton “Recurrent neural networks(RNN)”,handle?

Alternatively, aim of conducting this experiment is to learn how to curate datasets in order to train the “Seq-2-Seq” chatbot because there is a severe lack of education-based datasets for the program, particularly for the Malay language. In this study, we compared how well word and character embedding performed and the impacts of dropout [12] a Recurrent Neural Network variant on the quality of the response given by the Gated Recurrent Unit [13] (RNN). We looked into the “Seq-2-Seq”model's replies to various types of queries and its capacity to independently identify specific relevant Euphemisms are examples of data associations. We offer a thorough review of our research along with a report on a pilot experiment. In conclusion, this article's key contributions seem to be:
i)We include some evidence guidance on suitable “settings and optimizations” that researchers can utilize when designing a "Seq-2-Seq” model expressly towards the problem of questioning-answering. For instance, we found that a 4% failure degree increases the model's response capability..
ii) We provided examples of how to curate a specialized dataset to train a machine-learning algorithm. The model should be capable of learning specific information from the dataset rather than relying on pre-trained embedding or rules. This is helpful for special datasets, like the one we are using in Malay.
iii) A questioning-and-answering system should be capable of delivering (creating) answers to a variety of natural language question categories
There are five (5) sections in this essay. The introduction comes first (this section), then related works (section 2), where we briefly cover earlier experiments that we are aware of. The experimentalinfrastructure, which contains the dataset, assessment questioning, models assessed, &training software, is explained in section three (3). Experiment, Section four contains the results and an in-depth explanation. In section five, we wrap up this article.

2. Background study work:-

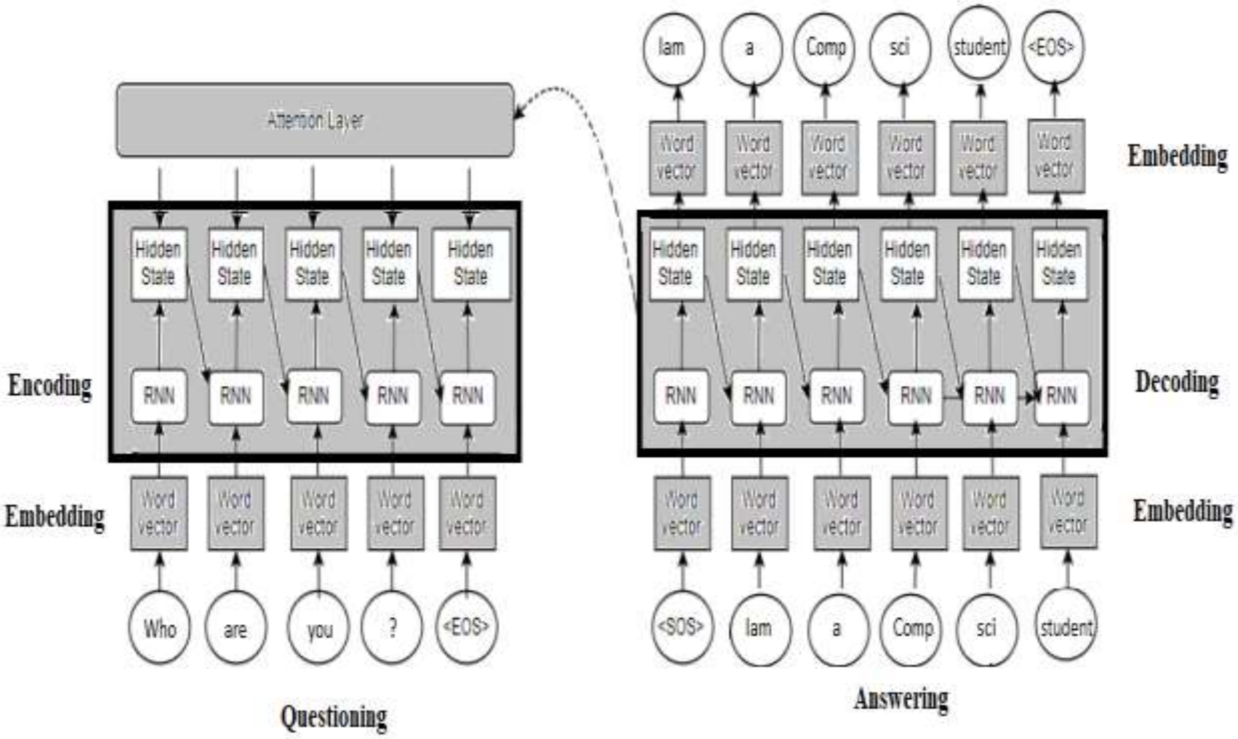
We were motivated to do this experiment by one of the most recent and nearby works, [14]. In order to test various settings and find some that worked well for a SeqSeq2 model, they carried out a number of experiments. As these settings are not relevant to our research objectives, we drew on their conclusions that bidirectional encoders are superior to unidirectional ones, beam searches of size 10 are ideal, and Bandana’s attention [15] mechanism outperformed Luong's attention [16] mechanisms. They did not, however, carry out any performance appraisal on the implications of varying failure rates (our RQ2). Additionally, we were incapable ofidentifyingany comparable trials that may address our RQ3 because it is so unusual and distinctive. For the most part.

2.1 Sequence to sequence model:-

Figure 1 depicts an example of A "Seq-2-Seq" model with word representations and an inductive method being used. The model is comprised of three (3) essential components:

i) Embedding - Embedding can be a word, a character, a bigram, a trigram, or a combination of the two. The embedding layer's task is to reshape the insert into something like a real-valued vector that can be utilized to depict it.
ii) Encoding: A bidirectional encoder is typically composed of a GRU network. The encoder's job is to convert the variable embedding'svectors used as input embedding, into intermediate states, which are fixed-length vectors.
iii) Decoding: The GRU network is used for decoding. The job of the decoder is to use beam search decoding to.

Fig. 1. Presentation of"Seq-2-Seq" model recognition



The "Seq-2-Seq" architecture has numerous components, as Observed in Figure 1 above, and each one can be used in a different way to address a particular problem or provide the desired results. For instance, despite the fact that word embedding was first employed for "Seq-2-Seq" models, it has significant limitations when applied to large datasets. The vocabulary will be enormous, requiring a lot of processing power and training time. Researchers set a restriction on the vocabulary size to overcome this issue. However, that in itself introduces a further constraint in that infrequently used words will be excluded from the lexicon, which will affect the model's learning and performance, particularly the question-answering system. The use of "character embedding" [17], [18], or "sub-word embedding" [19] is one way to get around this restriction. Using personality Failure rates should also be looked into in addition. For neural network-based models, dropout is a regularization method that was put forth by [20] to lessen over fitting. Randomly chosen neurons are ignored during training when dropout is used.As a result, other neurons are forced to fill in for the missing ones and gain knowledge of the internal network representations.

3 Experimentation:-

3.1 Dataset used in training:-

The pairs of hundreds of questioning and answering, the majority of which were given in Marathi language medium, were carefully curated into an experimentally small dataset in the applications of computer science and its domains. The database is formatted as a question, tab, and answer. There is one question and one answer per line in the file. The first five (5) question-answer pairings in the sample are displayed in Table 2.

Table 2.Shows some of the datasets of question-answer pairs.

Line. No	Question
1	पर्यवेक्षण न केलेल्या शिक्षणाचे उदाहरण द्या.(Grouping of the same question set)
2	पर्यवेक्षण न केलेल्या शिक्षणाचे उदाहरण द्या.(Grouping of the same question set)
3	पर्यवेक्षित शिक्षण म्हणजे काय? पर्यवेक्षित शिक्षणामध्ये, योग्य ग्रेड दिले जातात
4	पर्यवेक्षण न केलेले शिक्षण म्हणजे काय? संगणक स्वतःच त्याचा अंदाज बांधेल.
5	मजबुतीकरण शिक्षण म्हणजे काय? मजबुतीकरण शिक्षणामध्ये, खरे मूल्य नाहीदिले

The dataset's attributes are listed by as follows.

i) Cardinality mapping. The dataset comprises of mappings of questioning-answering pairs on both a one-to-one (1:1) and one-to-many (1:M) basis. The different mappings were added to the dataset to see how the different "Seq-2-Seq" model variations handled them, particularly one question to multiple responses mapping.

ii) Synonymous terms. To test if the model can literate-&-acknowledge the synonyms through the datasets, words with similar meanings are put in the dataset. These pairs (□□□□□□□□□□., □□□□□□□□□□.) and (neural, saraf) should be linked by the model as synonyms.

iii) Nearly identical inquiries or responses. To further complicate things, very similar questioning and/or answering pairsare been added to the datasets.

3.2 Question testing:-

In order to put our models to the test, we created two (2) categories of questions: Observed questions and unobserved questions. For more information, see Table 3

Table 3. Assessment of Questions

Classification of question	Question type to be assessed	No. of questions
Observed-1	1:1	8
Observed-2	1:M	2

Unobserved-1	Different word combinations	2
Unobserved-2	Substituted words (inside of vocabulary)	2
Unobserved-3	With spelling change (outside of vocabulary)	2
Unobserved-4	Addition in inquiry words (were found additionally in query)	2
Unobserved-5	Deletion in inquiry words (were missing in query)	2
Sum of		20

The likes of the world, of in the as in the, in One to one mapping, is a further division of Observed questions into those with only one correct response and those with multiple choices (one to many mappings). Unobserved questions are those that were not shown to the model during training. To comprehend how the "Seq-2-Seq" model responds to these, unobserved questions have been developed. The unobserved inquiries are realistic and genuine, reflecting the fact that different persons may pose the same topic in various ways. Various terms with similar meanings, different word orders, more complex or simpler question forms, some spelling changes brought on by linguistic familiarity (or unfamiliarity), or a simple error are examples of the differences (unintentionally). There are various unanswered questions.

i) Words in a different order - the questions' wording differs from that of the examples.

ii) Replaced words: A word from the Observed question has been swapped out for another word from the lexicon.

iii) Words with different spellings: One word in the Observed question has a different spelling (explicit from vocabulary word).

iv) Additional words in the questioning set - An observed question is given one additional word from the lexicon.

v) A shortened inquiry is asked by removing one word from the original one.

3.3 –Experimentation of the model:-

We tried four "Seq-2-Seq" model versions. Attrition degree and imbed categories (words or characters) vary between the models (Table 4).

Along with the variations mentioned above, the following fixed traits and hyperparameters were configured into each model:

i)GRU network bidirectional encoder

ii) The hidden and embedding sizes are 256 long.

iii) Word embedding for the replies (iii) (output)

iv) Instructional sampling size and initialization count are set to 10 & 200, respectively.

v) Size of the beam search (decoding) is 10.

The input and output vocabulary sizes for the word embedding models are 177 and 245 tokens, respectively. The input vocabulary size for the character embedding model versions is 47. They use the same output vocabulary of 245 items as the word embedding model. The token count for embedding includes computer-generated tokens such as "start" and "end."

4. Experimental results and discussion:-

4.1 Assessment criteria:-

The BLEU [21] rating is applied to assess the models. Bilingual Evaluation Understudy, or BLEU, was developed as an automated scoring system to compare translated and original material. In this example, we contrast the generated answer with the ideal response. A BLEU score ranges starting 0 to 1, where 0 indicates a complete lack of agreement and 1 indicates complete agreement, and anything in between indicates some degree of agreement between the two texts. Either answer is regarded as the right response for questions with a 1 to M mapping. The top models are those with the greatest BLEU scores.

4.2 Analysis and Results:-

Table 4 displays the total model outcome. Bold letters denote the superior version of each choice.

Table 4. Shows the model performance of overall (BLEU grading scores)

Failure	Embedded words	Embedded characters
0%	0.84	0.76
2%	0.94	0.70
4%	0.95	0.77
6%	0.09	0.61

RQ1-Which encoding forms best use keywords and characters for chatbots in education?
Outcome. Embedded word models consistently outperformed character embedding models for the same failure rates. In terms of the BLEU score, the best word embedding model outperformed the ideal character model, 0.95 versus 0.77.
Analysis. One of the disadvantages of "character-based models" is the length of the sequence. The sequence length of a character-based model for a similar sentence can be several orders of magnitude longer. This is a problem since character-based models could find it hard to represent long-distance dependencies. This is due to the fact that character-based models may make more predictions than word-based models.The potential for error increases as more predictions is made.For all versions, we solely used character embedding as input and word embedding as output to avoid this problem. Character model performance was, however, still inferior to that of word-based models.

RQ2-How does the performance of the education chatbot change when failure rates are applied?

Outcome. Applying various failure rates does have an impact on the model's performance, as observed in Table 4. At a failure rate of 40%, all models performed at their peak levels.
When no dropout or a failure rate of 60% was used, all models' performance slightly decreased.
Analysis. Without dropout, the models could get the right response for the known questions, but they struggled with the unknown questions, which may be a sign of over fitting. Over fitting decreased but performance increased, notably for the hidden questions when dropout was included.

RQ3-What kinds of education chatbots can the Sequence to Sequence model, which is based on recurrent neural networks, handle?

Table 5 Performance of model on the question classification (BLEU grading)

Classification of question	Embedded words for (4% failure)	Embedded charactersfor (4% failure)
Observed-1	1	1
Observed-2	1	0.93
Unobserved-1	1	1
Unobserved-2	1	0.01
Unobserved-3	1	0.05
Unobserved-4	0.05	0.05
Unobserved-5	1	0.61
Avg-Observed	**1**	**0.98**
Avg-Unobserved	**0.09**	**0.56**
Avg-Overall	**0.95**	**0.77**

This may be the experiment's central idea, making it an innovative experiment and valuable addition. The ability of the "Seq-2-Seq" models to handle Un-observed questions and questions with one question too many answers piqued our interest. The best performance ratings for each model variant are displayed in Table 5. The conclusions are as follows:-

i) The word-based model correctly responded to every question it was given (BLEU score: 1).

ii) Major disparity was discovered in the category of unanswered questions. Character-based models were only capable of scoring 0.5691 whereas word-based models were capable of getting 0.9.

iii) With the exception of questions with extra words, word-based models answered Un-observed questions accurately in all categories.

iv) Word models were capable of learning and correlating with synonym terms by properly responding to Un-observed questions that had these words switched around, such as (□□□□□□□□□□., with □□□□□□□□□□.) and neural with saraf.

v) It's also noteworthy to notice that one of the observed questions had a model that produced its own version of the right response (a questioning with one to many mappings). However it did not appear in the "training dataset", as shown in Table 6, the answer produced by the models was as follows: (□□□□□□□□□□., and □□□□□□□□□□.)mixed together in one sentence.

Table 6. Results of trained Synonyms

Trained synonyms	पर्यवेक्षण न केलेल्या शिक्षणाचे उदाहरण द्या.(Grouping and classification of the same question set)
	पर्यवेक्षण न केलेल्या शिक्षणाचे उदाहरण द्या.(Grouping and classification of the same question set)
	पर्यवेक्षित शिक्षण म्हणजे काय? पर्यवेक्षित शिक्षणामध्ये, योग्य ग्रेड दिले जातात
	पर्यवेक्षण न केलेले शिक्षण म्हणजे काय? संगणक स्वतःच त्याचा अंदाज बांधेल.
Asked questioned	मशीन लर्निंग अल्गोरिदमची श्रेणी सांगा
Replied questioned	पर्यवेक्षित आणि पर्यवेक्षी नसलेले मशीन लर्निंग

Analysis. This experiment has demonstrated to us the Sequence - to - sequence model's generative capability. It was capable of triggering an unanticipated response. Additionally, it should be noted that the "Seq-2-Seq" model can learn synonyms on its own if a suitable collection is provided (there is really no requirement for a supplementary set of guidelines, a lexicon. To make this a conclusive outcome, more study and testing are required.

5. Conclusion:-

We offer some suggestions for settings and enhancements that other researchers can use when creating a "seq-to – seq" model specifically for the "questioning-answering" issue in learning environments. Additionally, we provided a few examples of how a specific database can be selected to teach a model in order to allow it to learn specific information from the dataset without relying on norms or immovable embedding's that have already been trained. This is helpful for special databases, like, datasets in educational contexts that are in the Malay language. We have suggested an alternative.

Even though it was a modest trial, we undertook it because we wanted to learn more about a few specific variables and components of a straightforward Sequence - to - sequence model. Character embedding consistently performed worse than word embedding. We showed the Sequence - to - sequence model's power and the importance of training data for creating a model that performs well. Even while it might appear straightforward, adjusting failure rates showed a substantial enhancement in the prediction model, particularlyover lifting fitting and delivering reliable data without the need for additional complexities.

Natural language inquiry types that an automated system should be capable of supply (create) replies for.

6. References

[1] C. H. Lu, G. F. Chiou, M. Y. Day, C. S. Ong, and W. L. Hsu, "Using instant messaging to provide an intelligent learning environment," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4053 LNCS, pp. 575–583, 2006. https://doi.org/10.1007/11774303_57

[2] J. Jia, "CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning," Knowledge-Based Syst., vol. 22, no. 4, pp. 249–255, 2009. https://doi.org/10.1016/j.knosys.2008.09.001

[3] D. Griol and Z. Callejas, "An architecture to develop multimodal educative applications with chatbots," Int. J. Adv. Robot. Syst., vol. 10, 2013.

[4] D. Song, E. Y. Oh, and M. Rice, "Interacting with a conversational agent system for educational purposes in online courses," Proc. - 2017 10th Int. Conf. Hum. Syst. Interact. HSI 2017, pp. 78–82, 2017. https://doi.org/10.1109/hsi.2017.8005002

[5] S. Carayannopoulos, "Using chatbots to aid transition," Int. J. Inf. Learn. Technol., vol. 35, no. 2, pp. 118–129, 2018.

[6] D. Rooein, "Data-driven EDU chatbots," Web Conf. 2019 - Companion World Wide Web Conf. WWW 2019, pp. 46–49, 2019. https://doi.org/10.1145/3308560.3314191

[7] F. Clarizia, F. Colace, M. Lombardi, F. Pascale, and D. Santaniello, "Chatbot: An Education Support System for Student," in CSS 2018. Lecture Notes in Computer Science, vol. 1, Springer International Publishing, 2018, pp. 194–208. https://doi.org/10.1007/978-3-030- 01689-0_23

[8] R. Winkler and M. Soellner, "Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis," Acad. Manag. Proc., vol. 2018, no. 1, p. 15903, 2018. https://doi.org/10.5465/ambpp.2018.15903abstract

[9] L. Vygotsky, "The Development of Higher Psychological Processes," Mind Soc., vol. 6, no. 5, pp. 471–475, 1978.

[10] M. Alavi and D. E. Leidner, "Review : Knowledge Systems : Management Knowledge and Foundations Conceptual," MIS Q., vol. 25, no. 1, pp. 107–136, 2001. https://doi.org/10.2307/3250961

[11] O. Vinyals and Q. Le, "A Neural Conversational Model," Proc. 31st Int. Conf. Mach. Learn., vol. JMLR: W&CP, 2015.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

[13] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., pp. 1724– 1734, 2014. https://doi.org/10.3115/v1/d14-1179

[14] D. Britz, A. Goldie, M. Luong, and Q. Le, "Massive Exploration of Neural Machine Translation Architectures," Proc. ofthe 2017 Conf. Empir. Methods Nat. Lang. Process., 2017. https://doi.org/10.18653/v1/d17-1151.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR, pp. 1–15, 2015.

[16] M.-T. Luong, H. Pham, and C. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Proc. 2015 Conf. Empir. Methods Nat. Lang. Process., 2015. https://doi.org/10.18653/v1/d15-1166

[17] D. Golub and X. He, "Character-Level Question Answering with Attention," Proc. 2016 Conf. Empir. Methods Nat. Lang. Process., pp. 1598–1607, 2016.

[18] M.-T. Luong and C. D. Manning, "Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models," Proc. ofthe 54th Annu. Meet. ofthe Assoc. Comput. Linguist., 2016. https://doi.org/10.18653/v1/p16-1100

[19] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," ACL, 2016. https://doi.org/10.18653/v1/p16-1162

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," Proc. 40th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 311– 318, 2002. https://doi.org/10.3115/1073083.1073135