Cognitive Thinking: An International Journal of Interdisciplinary Studies
Volume-1, Issue-3 (July-September 2025), pp.128-134, ISSN: 3107-5088
www.cognitivethinking.in

A Comparison of RNN based Seq2Seq and Transformer Models for Detecting Similar Questions

1. Arwinder Singh
arwinder@pbi.ac.in
Department of Computer Science
University College, Ghanaur, Punjabi University, Patiala
2. Satpal Singh
spsingh.mohali@gmail.com
University College, Chunni Kalan, Punjabi University, Patiala

Abstract
There are millions and billions of questions on various knowledge sharing websites like Quora. It's very common that various questions are duplicate in it. An automatic tool for detecting similar questions out of these millions of questions is known as similar question detection approach. So, this paper presents an approach for detecting similar questions. The semantic similarity between pairs of questions represents that questions are similar. The approach detects similar questions by comparing the sentence embeddings. All the questions from Quora question pair dataset have been converted into sentence embeddings and then the given question is passed to the model for detecting similar questions. The comparison of detecting similar questions has been done with two methods i.e. Cosine and Jaccard similarity. The combination of Jaccard and cosine similarity can perform better for detecting similar questions as Jaccard similarity is able to detect lexical and cosine can detect semantic similarity. The important phase in this study is the conversion of questions into embeddings. So, the questions have been converted into embeddings using three models: average vector, RNN based Seq2Seq and transformer model. Two datasets have been used for the comparison: one is Quora question pairs dataset in English and second is the translated Quora question pairs in Punjabi. The BLEU similarity metric has been applied for the evaluation of the proposed similar question detection approach. The embeddings of the detected and the input question have also been compared for in-depth evaluation.
Keywords: Websites, Quora, Database, Lexical Similarity, Experiences.

1. Introduction

The similar question detection is a common problem, and it is very important in natural language processing. The task of similar question detection can also be seen as sentence similarity. The detection of similar questions is a very challenging task, and a lot of work is done on it. We have seen such problem on Quora where millions and billions of questions are duplicate. Users ask questions on this website and most of the time there are similar questions occur, and users answers those duplicate questions. So, it becomes very challenging for Quora to detect same questions.

An automatic tool is required for detecting similar questions and that tool is known as similar question detection approach. This paper presents an approach for detecting similar questions which takes the advantage of the current state-of-the-art transformer model. The approach detects similar questions by comparing the sentence embeddings. All the questions from Quora question pair dataset have been converted into sentence embeddings and then the given question is passed to the model for detecting similar questions. The important phase in this study is the conversion of questions into embeddings. So, the questions have been converted into embeddings using three models: average vector, RNN based Seq2Seq and transformer model, but the results of the average vector-based approach is very low as compared to others. So, the results using this metric has not been mentioned in result section. Two similarity measure metrics i.e. Cosine and Jaccard similarity have been applied for comparison between couple of questions. The combination of both metrics has been used to detect lexical as well as the semantic relatedness between pairs.

Two datasets have been used for the comparison: one is Quora question pairs dataset in English and second is the translated Quora question pairs in Punjabi. The BLEU similarity metric has been applied for the evaluation of the proposed similar question detection approach. The embeddings of the detected and the input question have also been compared for in-depth evaluation.

Problem Definition

The similar question detection problem is described in Figure 1, where two questions have been passed to the system and proposed system is required to detect whether the given pair of questions are similar or not. In a formal way,

$f(q_1, q_2) \rightarrow 0/1$ Eq. 1

In the above equation 1, f is a learned function. The output 0 means there is no similar question detected whereas 1 means similar question detected.

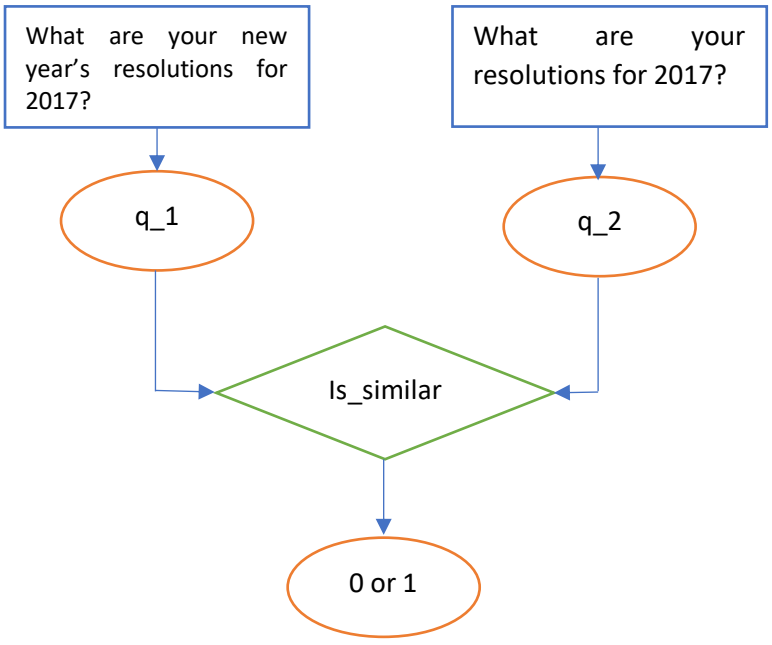

Figure 1: Problem formulation for detecting similar question

1. Related Work

A lot of research has been carried out for detecting similar questions. The machine learning [1,2] based models are very popular, and they proposed impressive results. The deep learning models helps the researchers for developing good NLP tools. The current research using deep

learning techniques has proved better results than traditional techniques. The detection of question similarity is a task for finding lexical as well as semantic similarity between pair of questions or sentences. The detection of lexical similarity is a straightforward and an easy task but to determine the semantic similarity is a challenging. The challenge for developing NLP based tools is the representation of large text on semantic space [3].
D. Bogdanova et al. [4] worked for encoding the input text into vectors using D-CNN (Deep Convolutional Neural Network). In their paper, the model converted input questions into fixed size vectors with Siamese architecture by sharing the parameters [5]. This study represents first time that can tackle duplicate questions. The authors in the paper presented two representations for input questions. Again, they used metrics for similarity calculation that can make comparisons between similar questions. The results of this paper show that CNN outperforms the existing machine learning models like SVM. Their study produced 92% of accuracy.

The work done by D. Bogdanova et al. [4] produced an accuracy of 73.4%. But the authors did more accurate in 2018 where J. Silva et al. [6] did something wrong where the model passed *duplicate* column for further training. And that increased the accuracy of the model up to 20%. The authors in J. Silva et al. [6] presented a new approach as hybrid model which is called *hybrid D-CNN*, it also produced 79% of accuracy. The dataset for their experiments was Ask-Ubuntu. The work in Y. Homma et al. [7] presented a Siamese-GRU/RNN system that is helpful to produce 84.9% of accuracy. The authors claimed that their model can handle 10 to 15 words.

Similar question detection is similar to paraphrase detection [8,9]. Paraphrase detection is a task for detecting that the given two sentences semantically similar or not. The work in [4] combined CNN and LSTM and the authors experimented on social media text for doing so. The most recent work implemented by [11,12,13,14,15]. A most recent work done on paraphrase detection by [16] in which the authors worked on paraphrase detection for Punjabi language. They developed a deep learning-based tool for the detection of paraphrases.

2. Dataset

A dataset of similar questions has been released by Quora in 2017, and it is commonly used for experiments for paraphrase generation & detection. This dataset has 400K similar questions. The paraphrasing pairs in this dataset has been marked with a numeric value 0 or 1. Here 0 represents that the questions are not similar and 1 means the questions are similar. This dataset is available in English, so this dataset has been used as such for question detection in English. Another experiment is done on Punjabi dataset. For this experiment, the English dataset needs to be translated into Punjabi. So, for the present work, this dataset was then translated into Punjabi language using Google translator.

Table 1: Statistics of Question Pairs collected from Quora

Item	**Value**
Question pairs	1,42,000
Total vocabulary	42,569

3. Pre-processing

The machine learning algorithms require noise free training data for better performance. Some of the sentences was not translated as accurate. So, we had to rectify the errors from that data. Some more pre-processing steps has to be done as follows:

1. Special characters removal
2. Remove too short and too lengthy sentences
3. Converting sentences into lower case
4. Tokenize the input questions

4. Experiments

The detection of similar questions among the millions of questions is a challenging as well as an important task. So, this article applies the advantage of current state-of-the-art transformer model for detecting similar questions. If we give a question q1, then detect a similar question q2 as shown in Figure 2 by comparing the Cosine and Jaccard similarity between their embeddings. Here, embeddings of English and Punjabi questions have been saved at different places, the input question in English will be compared with English questions whereas the input question of Punjabi will be compared with Punjabi questions. The embeddings of input data have been processed with RNN based Seq2Seq and transformer models. Different steps for detecting similar questions are shown in Figure 2.

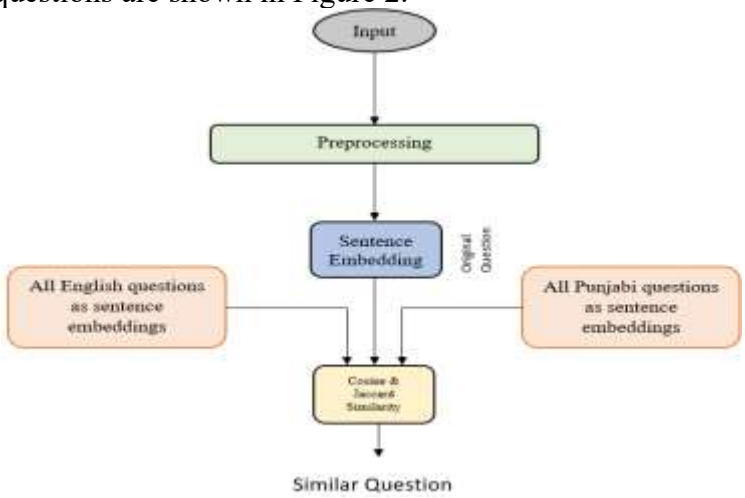

Figure 2: General Architecture of Similar Question Detection Method

Input

The questions in English & Punjabi will be the input to the above architecture.

1) Pre-processing

Removing of punctuations and special characters, remove too small and long questions and the last one is to tokenize the input strings are the various pre-processing steps performed on input questions.

2) Sentence Embedding

The next step it to convert an input question into embeddings, so transformer model has been applied here to do the same.

3) All Questions as Sentence Embeddings

There are three methods have been experimented for converting input data into embeddings. One method is RNN based Seq2Seq, second is transformer model to convert all questions into sentence embedding and the last one is average vectors approach. But the results of transformer are better than RNN based Seq2Seq and the results of average vector-based approach are very low. Two models have been trained separate for English as well as for Punjabi.

The input sentence of English type will then be compared with trained model of English data whereas Punjabi question will be compared with model trained of Punjabi dataset.

4) Cosine & Jaccard Similarity

The detection of similar questions is more dependent upon lexical as well semantic similarity. So, the combination of Cosine and Jaccard similarity metrics used for finding lexical as well semantic meanings. This comparison will give us the similar question as output.

6. Results and Discussions

A. Metrics

BLEU [17] is a well-known evaluation metric which has been proposed especially for translation models. But this metric has also been followed for evaluating the proposed similar question detection approach. This metric checks the lexical overlap between the input and output.

The output generated by the model developed in this paper then tested by comparing the sentence embeddings of the given and detected question. So, the comparison is made with cosine and combination of cosine and Jaccard similarity. The combination can detect lexical and semantic similarity between the pairs of sentences. The similarity using sentence embeddings is shown as Seq2Seq Sentence Embeddings Similarity (Seq2SeqSES) in Table 2. The results of BLEU and embeddings are shown in Table 2.

Table 2: Performance of the proposed model for Detecting Similar Questions

Similarity Metrics	**RNN based Seq2Seq Model on Punjabi Dataset**	**RNN based Seq2Seq Model on English Dataset**	**Transformer Model on English Dataset**	**Transformer Model on Punjabi Dataset**
BLEU_2	35.2	35.7	45.2	45.18
Jaccard Similarity	74.2%	73.1%	84.6%	83.2%
Cosine Similarity	71.5%	72.3%	72.1%	73.25%
Seq2SeqSES with cosine and Jaccard Similarity	81.8%	82.2%	87.3%	86.8%

The third evaluation metric is the qualitative method that has been applied for evaluation. Three experts in Punjabi selected and assigned them 500 questions of both English and Punjabi from the test set. The results of are shown in Table 3.

Table 3: Human evaluation For Detecting Similar Questions

Model	**Similar Questions**	**Partial Similar Questions**	**Dissimilar Questions**
RNN based Seq2Seq	83%	11%	6%
Transformer model	92%	5%	3%

The questions detected by the proposed study are shown in Table 4 where we can see that the system is able to detect similar questions.

Table 4: Examples of the Detected Similar Questions

		Input/Output
1	Input Question (Punjabi)	ਕੀ ਰੱਬ ਸੱਚਮੁੱਚ ਸੰਪੂਰਣ ਹੈ?

	Detected Question (Punjabi)	ਕੀ ਰੱਬ ਸੰਪੂਰਨ ਹੈ?
2	Input Question (Punjabi)	2017 ਲਈ ਤੁਹਾਡੇ ਨਵੇਂ ਸਾਲ ਦੇ ਮਤੇ ਕੀ ਹਨ?
	Detected Question (Punjabi)	2017 ਲਈ ਤੁਹਾਡੇ ਮਤੇ ਕੀ ਹਨ?
3	Input Question (English)	How can I become a good speaker?
	Detected Question (English)	How can I become a good public speaker?
4	Input Question (English)	How can I make money online for free?
	Detected Question (English)	How can I make money?

7. Conclusion

In the proposed work, a new approach for detecting similar questions has been developed for English as well as Punjabi language. The transformer model has been used for converting questions into embeddings for the approach developed in this paper. The questions have been converted into vectors/embeddings in the model that is presented in the proposed article. RNN based Seq2Seq and transformer model are used for representing the input data into embeddings or vectors. Two datasets have been used for the evaluation of the model presented in this article. One dataset is the Quora question pairs in English and second dataset is the translation of English pairs into Punjabi language. The BLEU similarity metric has been applied for the evaluation of the proposed similar question detection approach. The embeddings of the detected and the input question have also been compared for in-depth evaluation.

The architecture presented in this article is very effective and able to detect more accurate questions. The embeddings created with transformer model can detect semantic as well as syntactic which more important for finding semantic similarity between text. In future, the proposed architecture will be compared with BERT for improving the results.

Works Cited and Consulted

[1] Mishra, R., Barnwal, S.K., Malviya, S., Mishra, P., Tiwary, U.S.: Prosodic feature selection of personality traits for job interview performance. In: Abraham, A., Cherukuri, A.K., Melin, P., Gandhi, N. (eds.) ISDA 2018 2018. AISC, vol. 940, pp. 673–682. Springer, Cham (2020).

[2] Pathak, S., Ayush Sharma, S.S.S.: Semantic string similarity for quora question pairs (2019).

[3] Singh, S., Malviya, S., Mishra, R., Barnwal, S.K., Tiwary, U.S.: RNN based language generation models for a Hindi dialogue system. In: Tiwary, U.S., Chaudhury, S. (eds.) IHCI 2019. LNCS, vol. 11886, pp. 124–137. Springer, Cham (2020).

[4] Bogdanova, D., dos Santos, C., Barbosa, L., Zadrozny, B.: Detecting semantically equivalent questions in online user forums. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, pp. 123–131 (2015).

[5] Yu, J., et al.: Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 682–690 (2018).

[6] Silva, J., Rodrigues, J., Maraev, V., Saedi, C., Branco, A.: A 20% jump in duplicate question detection accuracy? Replicating IBM teams experiment and finding problems in its data preparation. META **20**(4k), 1k (2018).

[7] Homma, Y., Sy, S., Yeh, C.: Detecting duplicate questions with deep learning. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS) (2016).

[8] Vila, M., Mart´ı, M.A., Rodr´ıguez, H., et al.: Is this a paraphrase? What kind? Paraphrase boundaries and typology. Open J. Modern Linguist. **4**(01), 205 (2014).
[9] Zhu, W., Yao, T., Ni, J., Wei, B., Lu, Z.: Dependency-based siamese long shortterm memory network for learning sentence representations. PloS One **13**(3), e0193919 (2018).
[10] Agarwal, B., Ramampiaro, H., Langseth, H., Ruocco, M.: A deep network model for paraphrase detection in short text messages. Inf. Process. Manag. **54**, 922–937 (2018).
[11] El Desouki, M.I., Gomaa, W.H.: Exploring the recent trends of paraphrase detection. Int. J. Comput. Appl. **182**, 1–5 (2019).
[12] Yang, R., Zhang, J., Gao, X., Ji, F., Chen, H.: Simple and effective text matching with richer alignment features. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4699–4709. Association for Computational Linguistics, Florence, Italy (2019).
[13] Xu, S., Shen, X., Fukumoto, F., Li, J., Suzuki, Y., Nishizaki, H.: Paraphrase identification with Lexical, syntactic and sentential encodings. Appl. Sci. **10**, 4144 (2020).
[14] Yinfei, Y., Yuan, Z., Chris, T., Jason, B.: PAWS-X: a Cross-lingual Adversarial Dataset for Paraphrase Identification. CoRR, Abs/ **1908**(11828), 1–6 (2019).
[15] Mohamed, I., Hosam, W.: Exploring the recent trends of paraphrase detection. Int. J. Comput. Appl. **182**, 1–5 (2019).
[16] Singh, A. and Josan, G. S. 2021. A Deep Network Model for Paraphrase Detection in Punjabi. In Recent Innovations in Computing. Springer Singapore, 173-185.
[17] Papineni, K., Salim, R., Todd, W. and Wei J. Z. (2002), "BLEU: A Method for Automatic Evaluation of Machine Translation", In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, USA, pp. 311–318.