Chinese Medical Paraphrase Generation: Based on Neural Machine Translation

Bo Sun
Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Fei Zhang
Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Jing Yuan
Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Zhao Wei (zwfuwai@163.com)
Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College

Shu Ting
National Institute of Hospital Administration, National Health Commission

Research Article

Keywords: Paraphrase Generation, Machine Translation, Deep Text Matching Model, Deep learning

Posted Date: June 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-551021/v1

Chinese Medical Paraphrase Generation: Based on Neural Machine Translation

Bo Sun[1#] Fei Zhang[1#] Jing Yuan[1] Zhao Wei[1*^] Ting Shu[2*@]

[1] *Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, No.167 North Lishi Road, Xicheng District, 100037, Beijing, China.*

[2] *National Institute of Hospital Administration, National Health Commission, Building 3, yard 6, Shouti South Road, Haidian, 100044, Beijing, China.*

[#] *Bo Sun, Fei Zhang contributed equally to this work*

[*] *Zhao Wei, Ting Shu: Co-corresponding author*

[^] zwfuwai@163.com

[@] nctingting@126.com

Abstract：

Background: As people prefer to obtain medical knowledge online, medical intelligence question-answer systems based on question matching have attracted more and more attention, especially in China. However, due to the lack of paraphrase corpus of medical question, the development of this field is limited.

Objective: We propose a method for paraphrase generation which suitable for the Chinese medical field and use deep learning models instead of artificial evaluation for the first time. The method is designed to be able to automatically construct high quality Chinese medical paraphrase.

Methods: Validation experiments were carried out on two Chinese paraphrase data (one is general data, the other is medical data). Neural machine translation is used to generated paraphrase, that is, translate a sentence into other languages, and then reverse-translate it back to the original language to get the corresponding paraphrase. BLUE, ROUGEs, are used as quantitative evaluation metrics. Three deep text matching models are used to evaluate the generated paraphrase, instead of manual. Precision, Recall, F1 and AUC are used as qualitative evaluation metrics.

Results: 49908 and 4062 paraphrases were generated on the two datasets, and the generated efficiency was 97.03% and 98.38%, respectively. For the data in the two fields, the generated and original paraphrase pairs are very similar at the quantitative and qualitative evaluation metrics, especially the medical field. Take medical data as example, BLUE of generated and original paraphrase pairs are 0.556 and 0.626, respectively; the mean difference of AUC between the two groups was 0.015.

Conclusions: We first propose a paraphrase generation method based on neural machine translation and use deep text matching model instead of manual evaluation to evaluate the generated paraphrase. By analyzing the evaluation metrics，it can be concluded that：the paraphrase generated method has reached or even exceeded the level of artificial construction at the semantic level, especially in medical field; the deep text matching model can replace manual evaluation and realize automated paraphrase generation. This is of great significance to the development of Chinese medical paraphrase generation.

Keyword: *Paraphrase Generation, Machine Translation, Deep Text Matching Model, Deep learning*

Introduction

With the growth in the living standards, people are paying more and more attention to physical health, and hoping to obtain medical information conveniently online [1], especially during the covid-19 epidemic period. The famous online medical business websites in China include ‘haoyisheng.com’, ‘120ask.com’etc, while the well-known similar websites in other countries include ‘DailyStrength’, ‘MD-Junction’, etc. As times goes on, these websites’ question-answer record accumulate and form big data, which is the products of the wide participation of the people and contains a large number of real cases and high potential medical value [2]. With the constant

growth of medical data, we are all confronted with the problem of how to find answers to the questions we have [3]. Meanwhile, a large number of users—many of whom often ask similar, if not identical, questions—have placed a tremendous burden on the doctor-side and cause timely reply to be nearly impossible [4]. Thus, it is essential to develop techniques which can efficiently address the problem of medical question answering.

Question answering (QA) is an application of natural language processing (NLP) that tries to fulfill that need and has been receiving a lot of attention since the late 90s with evaluation campaigns such as TREC [5]. It is as pecialized type of information retrieval that returns precise short answers to queries posed as natural language questions [6-8]. The relevance and trustworthiness of the answers returned is of utmost importance in QA systems, and the latter especially for clinical domain [3]. Meanwhile, QA systems are often susceptible to the way questions are asked [9]. Thus, QA system based on question matching is getting more and more attention; namely, selecting automatically from some existing medical answer records the answer to the question that best matches a user's question. However, due to the lack of train data, the development of QA systems based on question matching has been greatly restricted. At present, there is a lack of large-scale similar question data, sspecially in the field of Chinese medicine. The problem can readily resolve by paraphrase generation task [10].

Paraphrases refer to texts with the same meaning but different expressions. For example, 'can "bailing capsule" be taken for a long time', 'can "bailing capsule" be taken for a long period' are paraphrases sentence pair. Paraphrase generation refers to a task in which given a sentence the system creates paraphrases of it [11]. Paraphrase generation is an important task in NLP, which can be a key technology in many applications, especially QA system.

Traditional, paraphrase generation has been addressed by using four methods, including: rule-based methods [12], thesaurus-based methods [13, 14], grammar-based methods [15] , statistical machine translation (SMT)-based methods [16, 17]. Recent advances in deep learning, in particular neural network based on sequence-to-sequence (Seq2Seq) learning, has made remarkable success in various NLP tasks, including machine translation [18], paraphrase generation [19, 20], etc. Zichao Li, et al [11], propose a new framework for the paraphrase generation, which consists of a generator and an evaluator, both of which are learned from data. Ankush Gupta, et al [10], proposed method is based on a combination of deep generative models (VAE) with Seq2Seq models (LSTM) to generate paraphrases, given an input sentence.

These studies focused on building a deep learning model in paraphrase generation task and achieved good results. However, deep learning model is supervised model, which means it's buliding need a lot of train materials (paraphrase data). In the field of Chinese medicine, there is a lack of paraphrase data which can be used for buliding.

In contrast to building new generation model, we propose to use mature neural machine translation (NMT) in paraphrase generation task. NMT is a Seq2Seq learning model for automated translation [18]. Compared with SMT, NMT has an overwhelming advantage, not only in the manual evaluation index, but also can reduce morphological errors, lexical errors and word order errors [21-23]. Meanwhile, the NMT has verified its performance in a real medical environment. Khoong, et al. [24] assessed the usefulness of machine translation in helping patients understand discharge instructions.

The another challenge in paraphrase generation lies in the definition of the evaluation measure [11]. Traditional, ROUGE, BLEU, etc. have been used as measure metrics, which could lose the calculating of semantic similarity. To quantify the aspects that are not addressed by automatic evaluation metrics, human evaluation becomes necessary. However, human evaluation will cost a lot of labor, and the results of evaluation could easily be subjectively affected. Hence, we propose to use the deep text matching models as an alternative to human evaluation. In recent years, with the development of NLP, a variety of matching models based on neural networks have been emerged and achieved good performance. The core of these models is similarity calculation, not only at the character level but also the sentence level [25-27].

In this study, we propose to use neural machine translation (NMT) in Chinese medical paraphrase generation task. It was verified on two Chinese paraphrase corpora, one is a general corpus, and the other is a medical corpus. BLEU and ROUGE metrics have been used in order to evaluate the results of approach. In addition, it is worth noting that we innovatively using deep text matching models instead of humans to evaluate the similarity.

Material and methods

Approach Description

The core of the approach is machine translation (MT). Literally understandable, machine translation is a technique that leverages computers to translate human languages automatically. MT models can be divided into two categories: statistical machine translation (SMT) and neural machine translation (NMT). NMT, which models direct mapping between source and target languages with

deep neural networks has achieved a big breakthrough in translation performance and and even approached human-level translation quality, especially parity on Chinese-to-English translation [28, 29]. At present, there are many translators based on neural machine translation, Google Translate (GT) [30]delivers roughly a 60% reduction in translation errors on several popular language pairs [18].

Based on the above, we aim to use NMT as generative model for Chinese medical paraphrase generation task. Specifically, it can be divided into two steps. The first step is using the Google Translate (GT) based on NMT to translate the Chinese original sentence into an English interlanguage sentence. In the second step, using the GT again to reversely translate the interlanguage sentence back to the Chinese form to obtain the paraphrase sentence. (Figure 1).

Data Sources

We evaluate our approach on two datasets, one of which (CCKS2018_Task) is a Non-medical field paraphrase dataset and the other (Chinese_covid) is a dataset in medical field.

CCKS2018_Task. It is the dataset of 'CCKS2018 WeBank Intelligent Customer Question Matching Contest'. This Contest is a real scene sentence intention matching task organized by the Intelligent Computing Research Center of Harbin Institute of Technology (Shenzhen). The dataset consists of 100K lines of Chinese question paraphrase pairs. Each line of data is composed of source sentence, reference sentence and labels. The label represents the similarity value of each paraphrase pairs, and the value is 0 or 1 (0 means dissimilar, 1 means similar).

Chinese_Covid. It is a Chinese medical dataset in 2020. The dataset consists of over 10K lines of question paraphrase pairs. Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the questions in the pair are truly similarily or not. Wherever the binary value is 1, the question pair is similarity.

The above data can be divided into pairs of similar and dissimilar paraphrase sentence pairs according to the label, that is, the label is 1 (denoted as positive question pairs) and the label is 0 (denoted as negative question pairs).

Evaluation

Quantitative Evaluation Metrics. For quantitative evaluation, we use the well-known automatic evaluation metrics: BLEU [31], ROUGE [32]. Previous work has shown that these metrics can perform well for the paraphrase recognition task [33] and correlate well with human judgments

in evaluating generated paraphrases [34]. Both of these scores lie between the range of 0 and 1 (or 0 and 100). Note that higher BLEU and ROUGE scores are better.

BLEU (Bilingual Evaluation Understudy) considers exact match between reference paraphrase(s) and generated paraphrase(s) using the concept of modified n-gram precision and brevity penalty. The score of it ranges from 0 to 1. The closer the score to 1, the higher the translation quality.

$$BLEU = \ BP \ \times \exp\left(\sum_{n=1}^{N} W_n \times \log Pn\right) \tag{1}$$

$$BP = \begin{cases} 1, & lc > lr \\ exp(1 - lr/lc), & lc \leq lr \end{cases} \tag{2}$$

Pn in the formula refers to the precision of N-gram. BP is the penalty factor. Wn refers to the weight of the N-gram, which is generally set as a uniform weight, that is, $Wn = 1/N$ for any n. lc: the length of the generated. lr: the length of the shortest reference.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) which is mainly based on the recall rate (Recall) including: ROUGE-N, ROUGE-L, etc.

$$ROUGUE - N = \ \frac{\sum_{S\in\{ReferenceSumamries\}} \sum_{gram_N \in S} Count_{match}(gram_N)}{\sum_{S\in\{ReferenceSumamries\}} \sum_{gram_N \in S} Count(gram_N)} \tag{3}$$

$$R_{LCS} = \ \frac{LCS(C,S)}{len(S)} \tag{4}$$

$$P_{LCS} = \ \frac{LCS(C,S)}{len(C)} \tag{5}$$

$$F_{LCS} = \ \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \ \beta^2 P_{LCS}} \tag{6}$$

ROUGE-N mainly counts the Recall on N-grams. The denominator of the formula is the number of N-grams in the reference, and the numerator is the number of N-grams shared by reference and genereated. ROUGE-L uses the longest common subsequence of generated C and reference S when calculating, L is the longest common subsequence (LCS). R_{LCS}in the formula is the recall rate, while $P_{LCS}$is the accuracy, F_{LCS} is ROUGE-L.

Qualitative Evaluation Metrics. To quantify the aspects that are not addressed by automatic evaluation metrics, human evaluation becomes necessary. However, human evaluation will cost a lot of labor, and the results of evaluation could easily be subjectively affected. Hence, we propose to use the deep text matching models as an alternative. The main steps of this qualitative evaluation method can be summarized as follow (Figure 2): First, the original question pairs (source

and reference) in each data source are split into train, valid and test set according to the ratio of 8:1:1. The test set here is also called the original test set. Second, combining the source sentence in the original test set with its corresponding generated sentence to form generated test set. Third, using train and valid set to bulid deep text matching models. In this study, three depth matching models were selected, including: KNRM, MVLSTM and Pyramid. Forth, using the original and generated test set to inference on the three trained models. Each model returns a number between 0 and 1 for each question pairs, which indicates the probability that whether the question pair is semantically similar or not. Precision, Recall, F1and C-statistic are used to compare the performance of each model on the two test sets. Precision measures accuracy; Recall measures comprehensiveness; F1 considers both comprehensiveness and accuracy; AUC assessed discrimination. The more similar the performance of the two test sets on these four metrics, the higher the performance of this method.

Deep text matching model

In this paper, we used deep text matching models (K-NRM [35]，MVLSTM [26] and Pyramid [27]) as an alternative to evaluate the similarity between source sentence and generated sentence. The steps of calculating the semantic similarity of the three models are basically the same, which can be divided into three steps: word representation, feature extraction and multi-layer perception. ***Word Representation***. Computer can't directly process the sentence. Words need to be represented in a vector or matrix first. Usually, all words in the sentence are represented by a fixed length word vector respectively, which called word embedding, such as Word2Vec[36], Glove[37],etc. ***Feature extraction layer***. For the results obtained by word embedding, methods such as matching matrix, are used to perform data dimensionality reduction on the basis of retaining the semantic features of the two sentences, and output feature vectors of the two sentences. ***Multi-Layer Perception***. For the feature vector obtained above, use a MLP (Multi-Layer Perception) to produce the final matching score.

K-NRM. It is a kernel based neural model for document ranking. Given a query and a set of documents, it uses a translation matrix that models word-level similarities via word embeddings, a new kernel-pooling technique that uses kernels to extract multilevel soft match features, and a learning-to-rank layer that combines those features into the final ranking score. The whole model is trained end-to-end. The kernels transfer the feature patterns into soft-match targets at each similarity level and enforce them on the translation matrix.

Pyramid. Inspired by the success of convolutional neural network in image recognition, where neurons can capture many complicated patterns based on the extracted elementary visual patterns such as oriented edges and corners, this method propose to model text matching as the problem of image recognition.

MVLSTM. This method presents a new deep architecture to match two sentences with multiple positional sentence representations. Specifically, each positional sentence representation is a sentence representation at this position, generated by a bidirectional long short term memory (Bi-LSTM). The matching score is finally produced by aggregating interactions between these different positional sentence representations, through k-Max pooling and a multi-layer perceptron.

Results

Overview of generated sentences

For the method proposed in this article, we have verified it on the two datasets CCKS2018_Task and Chinese_Covid. Specifically, for the source sentences of the positive question pairs (label = 1) in the two databases, using a NMT (in this article, it is Google Translate) to translate it into an interlanguage form (English), and then inversely translate it back to the original language form (Chinese) to construct paraphrase sentence (denoted as generated sentence). The results are as follows.

CCKS2018_Task3 data set contains 10W question pairs, of which 49,908 are positive sample data (sentence similarity = 1). After two translations by Google Translator (Chinese-English, English-Chinese), there are 1484 generated sentences and source sentences exactly the same. After excluding them, we finally constructed 48,424 valid paraphrase question pairs, with an effective ratio of 97.03%. Table 2.

Chinese_Covid contains 1W question pairs, of which 4062 are positive sample data (sentence similarity = 1). After the same process, 3996 valid paraphrase question pairs, were finally constructed, with an effective ratio of 98.38%. Table 3.

Evaluation of the Performance

We conducted experiments on the aforementioned data set and reported the qualitative and quantitative results of our method. The quantitative results for CCKS2018_Trask and Chinese_Covid datasets are given in Tables 4. For the CCKS2018_Trask data set, the BLEU value

of original question pairs is 0.244, however, the BLUE value of generated question pairs, which constructed by this method, is up to 0.412. Other evaluation metrics also have the same trend. The calculated result on the generated question pairs is slightly larger than that on the original question pairs. For the Chinese_Covid data set, the calculated result on the generated question pairs (ROUGE-1: 0.605) is slightly lowever than that on the original question pairs (ROUGE-1: 0.678).

Three deep text matching models are bulit with the train and valid set and tested with original test set and generated test set. In Tables 5 and Figure 4, we report the qualitative results from various models for the CCKS2018_Task and Chinese_Covid datasets respectively.

The average F1 values of original test set and generated test set in CCKS2018_Task are **0.826** and **0.836** respectively. The average Precision values of original test set and generated test set in Chinese_Covid are **0.806** and **0.811** respectively. In addition, we made ROC curve and C- statistic to show whether the performance of these models is different between the two test sets or not. The green line in the Figure 4 represents the ROC curve of the original test set, and the purple line represents generated test set. Taking Chinese_Covid as the example, the ROC curves of two test sets basically overlap each other, the AUC values are similar, and the maximum difference is 0.02.

Conclusions

In this paper, we first propose a method of Chinese medical paraphrase generation based on neural machine translation and use deep text matching model instead of manual evaluation to evaluate the generated paraphrase. Validation experiments were carried out on two Chinese paraphrase data. By analyzing the evaluation indicators such as BLUE, ROUGE and AUC, it can be concluded that： The paraphrase generated method has reached or even exceeded the level of artificial construction at the semantic level,especially in medical field; Deep text matching models can replace manual evaluation and realize automated paraphrase corpus construction. This is of great significance to the development of this field. This method can quickly and automatically construct a high-quality medical paraphrase corpus. It is helpful to promote the development of medical intelligent question answering system based on question matching.

Disscussion

In this study, we propose a method of medical paraphase generation，which based on NMT and verify its performance in Chinese test set. In view of the shortcomings of the current research on the generated methods of paraphrase, this study mainly makes the following contributions：

Using mature NMT for paraphrase generation without training data. At present, most of the paraphrase generation models are end-to-end models based on deep learning. In the process of model building, a large amount of high-quality paraphrase corpus is needed for trainin. In the field of Chinese medicine, such data is lacking. In view of this situation, we propose for the first time to use NMT (GT in this study) as a substitute for paraphrase generation models to generate paraphrase. The quantity of generated paraphrases was determined using the standard similarity[38]-BLEU and ROUGEs. METEOR (Metric for Evaluation of Translation with Explicit ORdering)[39], a commonly used standard similarity, uses WordNet to calculate the matching relationship of synonyms while calculating the similarity. This article studies the field of Chinese medicine, however, WordNet[40] is an english electronic lexical database, so METEOR is discarded.

From Table 5, we can know that for the two test sets, the test results of the original question pairs and the genetated question pairs of each standard similarity are similar. This shows that the paraphrase generated method presented has reached the level of artificial construction in the quantitative evaluation. Specifically, for CCKS2018_Task, results of the generated question pairs are slightly larger than the original question pairs. For Chinese_Covid, the results are opposite. BLEU and ROUGEs are evaluated from the word level, especially ROUGEs include ROUGE-1, ROUGE-2, ROUGE-L. From the above formula (3-6), ROUGE-1 counts the Recall on 1-grams，ROUGE-2 counts the Recall on 2-grams，ROUGE-L uses the longest common subsequence of generated and reference sentence when calculating. At present, the translation accuracy of GT for medical words is not very well[41, 42]. Therefore, for the medical set - Chinese_Covid, the generated question pairs (BLUE: 0.556; ROUGE-1: 0.605) constructed by this method are lower than the original question pairs (BLUE: 0.626; ROUGE-1: 0.678) in quantitative evaluation.

We present a brief review of the existing work. Unfortunately, there is no study on the Chinese medical paraphrase generation. So, we compared our model with other three related English medical paraphrase generation studies (Table 6) in BLUE. The BLUE of our model are higher than that

reported for most of the paraphrase generation models in the literature, indicating that our model had better prediction performance.

Using deep text matching model instead of manual evaluation. In the current study, manual evaluation is the core qualitative evaluation of paraphrase corpus, which leads to the failure of automatic generation. The evaluation results are easily affected by the subjective influence of evaluator. This has affected the development of this field. We propose to use the mature deep text matching model instead of manual. Text similarity calculation is the core of deep text matching model, which can evaluate the similarity of two sentences at the semantic level. Based on this, we choosed three representative deep text matching models as qualitative metrics.

We used Precision, Recall and F1 to evaluate whether there are differences between the two tests' (original test set and generated test set) results or not. After verification, the results calculated by the deep text matching model are consistent with those calculated by quantitative metrics: BLIUE and ROUGE. On the Chinese_Covid, results of the generated test set are better than the original test set, and the gap is very samll. Gap of Precision, Recall, F1 and AUC are: 0.005, 0.011, 0.06, 0.015(mean). The small gap is also reflected in the quantitative evaluation metrics. Gap of BLUE, ROUGE-1 are: 0.007 and 0.073.

In particular, from figrue 4, we can see that the ROC curves of the three models on the original and generated test set are very similar, and the difference in AUC values is very small: 0.015 (mean). From the above, training and valid set segmented by CCKS2018_Task and Chinese_Covid are used to bulid the text matching model, respectively. We think that the built models have learned the semantic information. Original and generated Test Set are tested in these models, to judge whether the paraphrase constructed by the method presented can reach the artificial paraphrase data at the semantic level. The results show that it has been reached. The deep text matching model, which considers more semantic information, is not sensitive to wrong vocabulary, comparing with traditional quantitative metrics. These explains that Chinese-Covid's generated question pairs are lower than original question pairs in quantitative metrics, and generated test set are higher than original test set in qualitative metrics.

There are also some defects in this study, such as: only used GT as neural machine translator; inaccurate translation of medical terms. In future study, we will use other neural machine translators and GT for comparative experiments and use the combination of medical term vocabulary

(SNOMED-CT, etc) and named entity recognition (NER), which is one of Natural language processing (NLP) technology, to solve the problem of medical term translation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

NO applicable.

Authors' contributions

WZ, TS and BS conceived the study and developed algorithm.

FZ collected and preprocessed the data.

YJ and BS designe dexperimental and result analysis.

BS and FZ carried out all the experimentand wrote the first draft of the manuscript.

All authors have read and approved the manuscript.

Acknowledgements

Not applicable.

Author details

[1]Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, No.167 North Lishi Road, Xicheng District, 100037, Beijing, China. [2]National Institute of Hospital Administration,

National Health Commission, Building 3, yard 6, Shouti South Road, Haidian, 100044, Beijing, China.

References

1. Liu X, Zhou Y, Wang Z: Deep neural network-based recognition of entities in Chinese online medical inquiry texts. *Future Generation Computer Systems* 2020, 114.
2. Swar B, Hameed T, Reychav I: Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. *Computers in Human Behavior* 2017, 70(MAY):416-425.
3. Cruchet S, Boyer C, Plas LVD: Trustworthiness and relevance in web-based clinical question answering. *Studies in Health Technology & Informatics* 2012, 180(1):863.
4. Sheng Z, Xin Z, Hui W, Cheng J, Pei L, Ding Z: Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs. *Applied Sciences* 2017, 7(8):767-.
5. Voorhees E: The TREC-8 Question Answering Track Report. In: *Text Retrieval Conference: 1999*; 1999.
6. Lee M, Cimino J, Hai RZ, Sable C, Hong Y: Beyond Information Retrieval—Medical Question Answering. *Amia Annu Symp Proc* 2006:469-473.
7. Cohen, Michael A, Hersh, William R: A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 2005.
8. Athenikos SJ, Han H: Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine* 2010, 99(1):1-24.
9. Fader A, Zettlemoyer L, Etzioni O: Open question answering over curated and extracted knowledge bases. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* New York, New York, USA: Association for Computing Machinery; 2014: 1156 - 1165.
10. Gupta A, Agarwal A, Singh P, Rai P: A Deep Generative Framework for Paraphrase Generation. 2017.
11. Li Z, Xin J, Shang L, Hang L: Paraphrase Generation with Deep Reinforcement Learning. 2017.
12. McKeown K: Paraphrasing Using Given and New Information in a Question-Answer System. *Technical Reports (CIS)* 1979.
13. Bolshakov IA, Gelbukh A: Synonymous Paraphrasing Using WordNet and Internet. In: *International Conference on Application of Natural Language to Information Systems: 2004*; 2004.
14. Kauchak D, Barzilay R: Paraphrasing for automatic evaluation. In: *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA: 2006*; 2006.
15. Narayan S, Reddy S, Cohen SB: Paraphrase Generation from Latent-Variable PCFGs for Semantic Parsing. In: *International Conference on Natural Language Generation.* 2016: 153-162.
16. Quirk C, Brockett C, Dolan WB: Monolingual Machine Translation for Paraphrase Generation. In: *Conference on Empirical Methods in Natural Language Processing: 2004*; 2004.
17. Zhao S, Cheng N, Ming Z, Liu T, Sheng L: Combining Multiple Resources to Improve SMT-based Paraphrasing Model. In: *Acl, Meeting of the Association for Computational Linguistics, June, Columbus, Ohio, Usa: 2008*; 2008.

18. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K *et al*: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016.
19. Cao Z, Luo C, Li W, Li S: Joint Copying and Restricted Generation for Paraphrase. 2016.
20. Su Y, Yan X: Cross-domain Semantic Parsing via Paraphrasing. 2017.
21. Bentivogli L, Bisazza A, Cettolo M, Federico M: Neural versus Phrase-Based Machine Translation Quality: a Case Study. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: 2016*; 2016.
22. Bojar O, Buck C, Federmann C, Haddow B, Tamchyna A: Findings of the 2014 Workshop on Statistical Machine Translation. In: *Workshop on Statistical Machine Translation: 2014*; 2014.
23. Junczys-Dowmunt M, Dwojak T, Hoang H: Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In: *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT): 2016*; 2016.
24. Khoong EC, Alicia F: Unsound Evaluations of Medical Machine Translation Risk Patient Health and Confidentiality—Reply. *Jama Internal Medicine* 2019, 179(7:4):1001.
25. Wan S, Lan Y, Xu J, Guo J, Pang L, Cheng X: Match-SRNN: modeling the recursive matching structure with spatial RNN. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence.* New York, New York, USA: AAAI Press; 2016: 2922–2928.
26. Wan S, Lan Y, Guo J, Xu J, Pang L, Cheng X: A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. 2015.
27. Pang L, Lan Y, Guo J, Xu J, Wan S, Cheng X: Text Matching as Image Recognition. 2016.
28. Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huang X, Junczys-Dowmunt M, Lewis W, Li M *et al*: Achieving Human Parity on Automatic Chinese to English News Translation. 2018.
29. Zhang J, Zong C: Neural Machine Translation: Challenges, Progress and Future; 2020.
30. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016.
31. Papineni K: BLEU : a method for automatic evaluation of MT. *Research Report, Computer Science RC22176 (W0109-022)* 2001.
32. Lin C-Y: ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out: jul 2004; Barcelona, Spain*: Association for Computational Linguistics; 2004: 74-81.
33. Madnani N, Heilman M, Tetreault J, Chodorow M: Identifying High-Level Organizational Elements in Argumentative Discourse. 2012.
34. Wubben S, Bosch Avd, Krahmer EJ: Paraphrasing Headlines by Machine Translation Sentential Paraphrase Acquisition and Generation using Google News. *lot occasional series* 2011, 16:169-183.
35. Xiong C, Dai Z, Callan J, Liu Z, Power R: End-to-End Neural Ad-hoc Ranking with Kernel Pooling. 2017.
36. Mikolov T, Chen K, Corrado G, Dean J: Efficient Estimation of Word Representations in Vector Space. *Computer Science* 2013.
37. Pennington J, Socher R, Manning C: Glove: Global Vectors for Word Representation. In: *Conference on Empirical Methods in Natural Language Processing: 2014*; 2014.
38. Madnani N, Tetreault J, Chodorow M: Re-examining Machine Translation Metrics for Paraphrase Identification. In: *Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies: 2012; 2012.
39. Lavie A, Agarwal A: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. 2007.
40. Fellbaum C: WordNet: An Electronic Lexical Database. *library quarterly information community policy* 1998.
41. Moberly T: Doctors are cautioned against using Google Translate in consultations. *BMJ (online)* 2018, 363.
42. Chen X, Acosta S, Barry AE: Evaluating the Accuracy of Google Translate for Diabetes Education Material. *JMIR Diabetes* 2016, 1(1).
43. van den Bercken L, Sips RJ, Lofi C: Evaluating Neural Text Simplification in the Medical Domain. In: *Web Conference 2019: Proceedings of the World Wide Web Conference.* New York: Assoc Computing Machinery; 2019: 3286-3292.
44. Soni S, Roberts K: Paraphrasing to improve the performance of Electronic Health Records Question Answering. 2020.
45. Adduru V, Hasan SA, Liu J, Yuan L, Datla V: Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In: *The 3rd International Workshop on Knowledge Discovery in Healthcare Data: 2018*; 2018.

Figures

Figure 1: Illustration of paraphrase generation based on neural machine translation (NMT). Neural machine translation: It is an Seq2Seq model following an encoder-decoder framework that usually includes two neural networks respectively

Figure 2: The main steps of qualitative evaluation method

Figure 3: The illustration of similarity calculation.

Figure 4: ROC curves of CCKS2018_Task and Chinese_Covid. ROC: receiver operating characteristic curve

Tables

Table 1: Description of data sources

Data sources	Number, n	Label	Domain	Label = 1, n
CCKS2018_Task	100000	0/1	General	49908
Chinese_Covid	10000	0/1	Medical	4062

Table 2: Examples paraphrases generated on CCKS2018_Task

Source	怎么今天登录不了？
Reference	无法登录

Generated	我为什么今天不能登录？
Source	怎么更换款卡
Reference	还款卡片有改动怎么办？
Generated	如何更改款卡

Table 3: Examples paraphrases generated on Chinese_Covid

Source	急性大咯血的症状有哪些？
Reference	急性大咯血会有什么症状呢？
Generated	急性大咯血的症状是什么？
Source	小儿支原体肺炎的复查咨询
Reference	小儿支原体肺炎有必要复查吗？
Generated	小儿支原体肺炎的复查和咨询

Table 4: Results of BLUE, ROUGE on CCKS2018_Task and Chinese_Covid dataset.

Data Source	BLUE	ROUGE-1	ROUGE-2	ROUGE-L
CCKS2018_Task:				
Original question pairs	0.244	0.299	0.085	0.287
Generated question pairs	0.412	0.450	0.172	0.428
Chinese_Covid:				
Original question pairs	0.626	0.678	0.436	0.640
Generated question pairs	0.556	0.605	0.318	0.564

Table 5: Results of Precision, Recall and F1 on CCKS2018_Task and Chinese_Covid dataset.

Data Source	Deep match model	Original Test Set			Generated Test Set		
		Precision	Recall	F1	Precision	Recall	F1
CCKS2018_Task	K-NRM	0.781	0.837	0.792	0.824	0.922	0.839
	Pyramid	0.818	0.830	0.820	0.845	0.883	0.851

	MVLSTM	0.880	0.870	0.878	0.841	0.791	0.832
	Average	**0.826**	**0.845**	**0.830**	**0.836**	**0.865**	**0.840**
Chinese_Covid	K-NRM	0.792	0.821	0.760	0.800	0.843	0.772
	Pyramid	0.824	0.803	0.786	0.838	0.832	0.805
	MVLSTM	0.802	0.771	0.758	0.794	0.751	0.745
	Average	**0.806**	**0.798**	**0.768**	**0.811**	**0.809**	**0.774**

Table 6: Comparison paraphrase generation of models in the literature

Study	Data		BLUE
	Name	Number	
Our stduy	Chinese_covid	10,000	0.556
Van et al.[43]	Expert+ Automated	6,064	0.548
Soni et al.[44]	CLINIQPARA	10,000	0.333
Adduru et al.[45]	WikiSWiki	1491	0.099

Figures

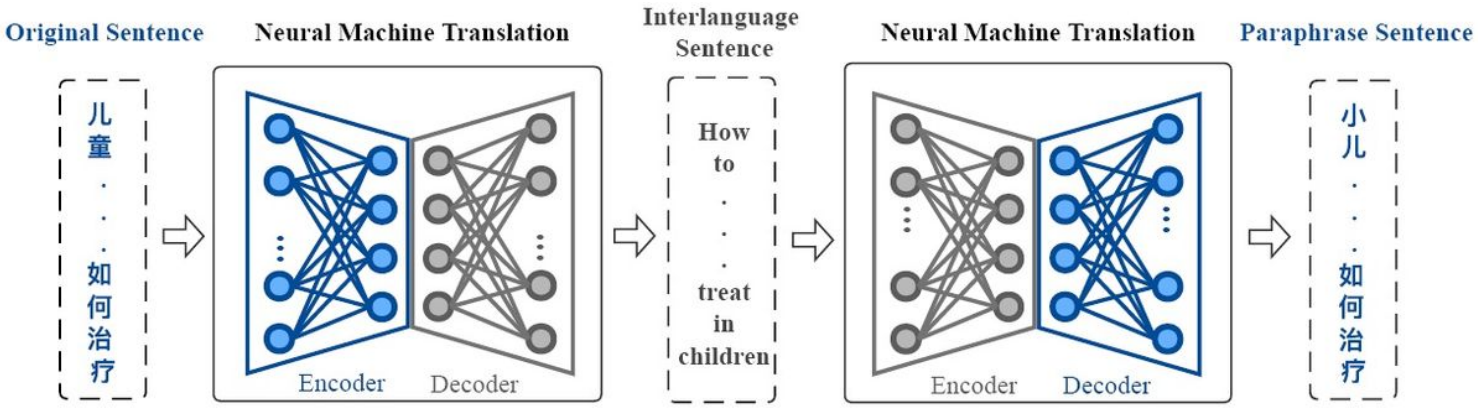

Figure 1

Illustration of paraphrase generation based on neural machine translation (NMT). Neural machine translation: It is an Seq2Seq model following an encoder-decoder framework that usually includes two neural networks respectively

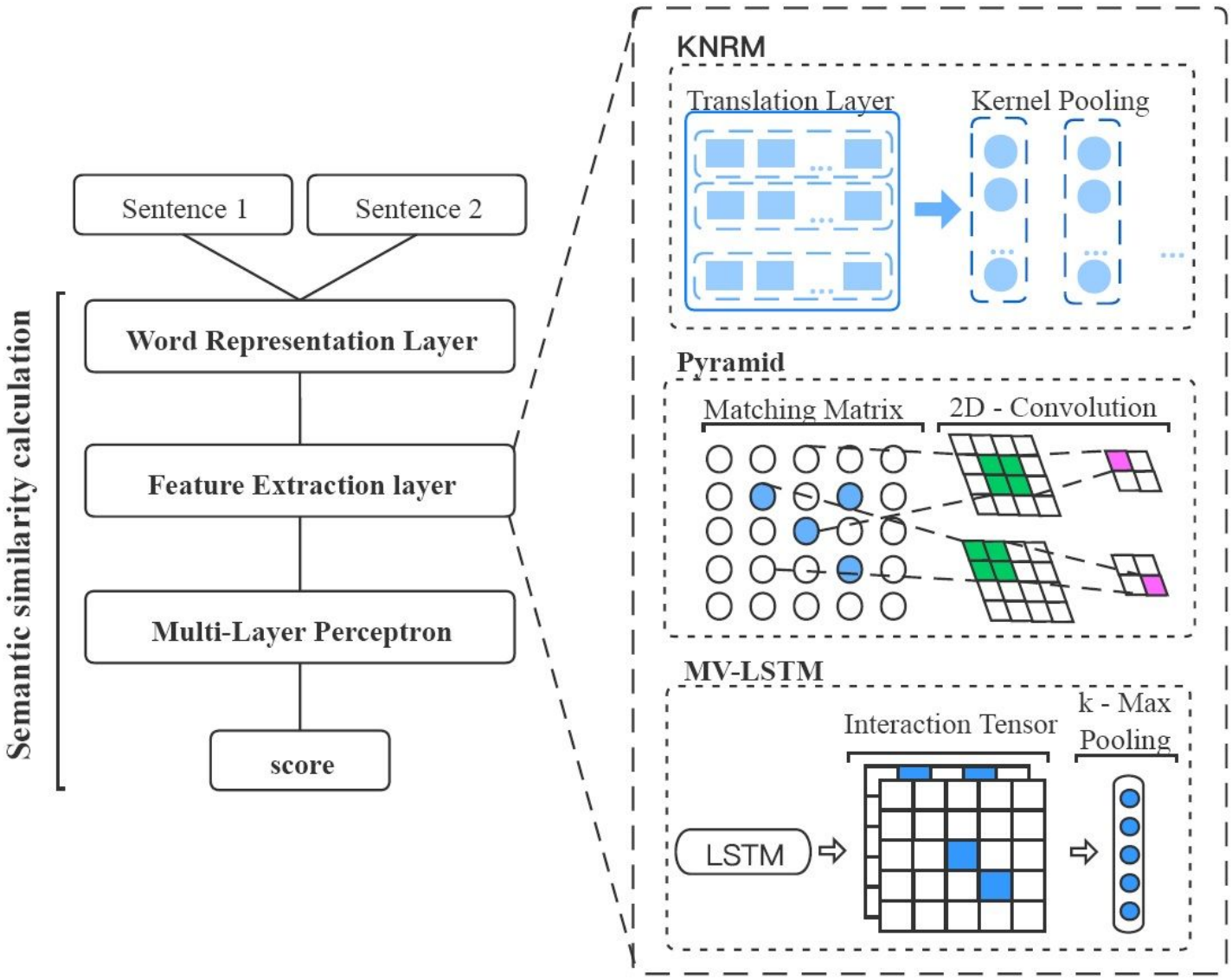

Figure 2

The main steps of qualitative evaluation method

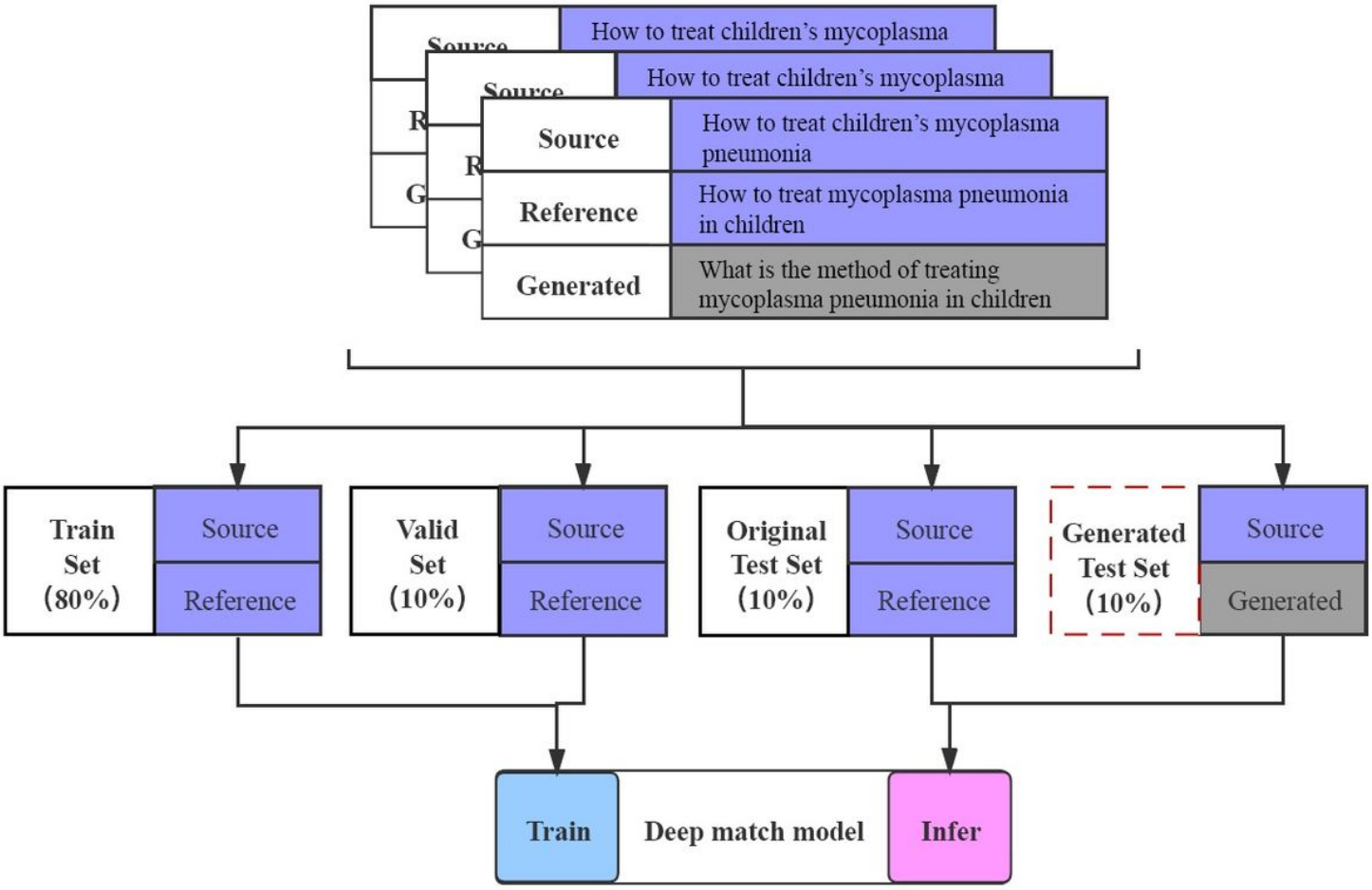

Figure 3

The illustration of similarity calculation.

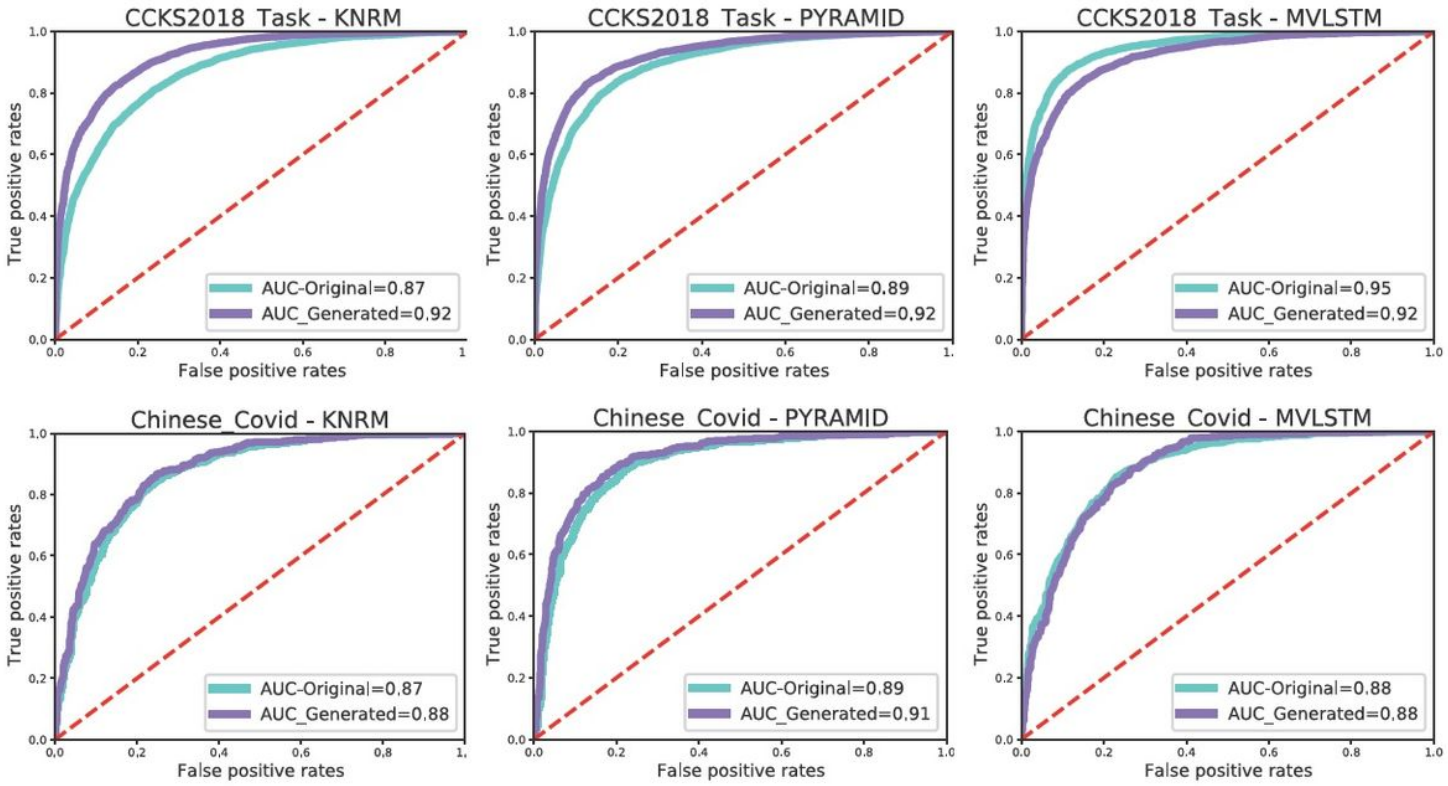

Figure 4

ROC curves of CCKS2018_Task and Chinese_Covid. ROC: receiver operating characteristic curve