2020 International Conference on Computational Science and Computational Intelligence (CSCI) | 978-1-7281-7624-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSCI51800.2020.00113

A Summary Evaluation Method Combining Linguistic Quality and Semantic Similarity

Xingwen Wang, Bo Liu*, Libin Shen, Yong Li
School of Software Engineering, Faculty of Information Technology
Beijing University of Technology
Beijing 100124, China
xingwen_w@126.com, boliu@bjut.edu.cn,
shenlevi@163.com, li.yong@bjut.edu.cn

Rentao Gu
Beijing Laboratory of Advanced Information Networks,
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing 100876, China
rentaogu@bupt.edu.cn

Guangzhi Qu
Computer Science and Engineering Dept.
Oakland University
Rochester, MI 48309
qu@oakland.edu

***Abstract*—Summary evaluation method is crucial to promote the development of text summarization technologies. However, most of the existing summary evaluation methods seldom consider the content integrity and readability of summaries simultaneously. This paper proposed a Linguistic Quality and Semantic Similarity Model (LQSSM) to evaluate generated summaries more comprehensively. Considering the readability of summaries, a linguistic quality evaluation network (LQEN) is proposed to evaluate summaries automatically. Meanwhile, a semantic similarity evaluation network (SSEN) is introduced to directly measure the informativeness between the summary and original text. Additionally, there is also a comprehensive evaluation using fusion indicators. The LQSSM and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) methods were used to score the summaries generated by five models. The results showed that the proposed method gave further feedback from different angles on the basis of reaching the level of other methods.**

Keywords- text summarization, summary evaluation, BERT, linguistic quality, semantic similarity

I. INTRODUCTION

The explosive growth of the Internet has result in information overload. Obtaining key information from massive texts is necessary. Text summarization, as a major research direction in the field of natural language processing, can extract content that conforms to the central idea from texts and organize it into a summary [1]. Text summarization is closely related to the summary evaluation method which can bring feedback to the quality of the text summary, reflect the shortcomings of the text summary technology, and promote the development of text summary technology.

The evaluation of the text summarization can be divided into manual evaluation and automatic evaluation. Manual evaluation is evaluated manually by experts and the results are reliable but the cost of time is relatively expensive, therefore, it is necessary to research the automatic evaluation methods. Automatic evaluation methods can be classified into two categories [2]: The first is intrinsic evaluation, which directly analyses the quality of the summary, it mainly evaluates the coherence and content integrity of the summary [3]. The second is extrinsic evaluation, evaluates the summary according to how it affects the completion of other tasks such as classification and indexing. Among them, extrinsic evaluation is more limited, and intrinsic evaluation is the mainstream summary evaluation method due to it can be applied to an independent environment. This paper mainly focuses on intrinsic evaluation.

The current intrinsic evaluation methods are mainly based on the comparison between system summary and human summary, but this will omit certain information, such as the inability to count the newly generated words. Besides, the coherence of summary and whether the expression is reasonable can only be evaluated by the subjective feelings of the evaluator. Thus, it's also a problem that it cannot automatically evaluate the linguistic quality of summaries. To address these problems, we proposed a solution mentioned in the paper.

The main contributions of our paper are summarized as below:

- This paper proposed Linguistic Quality and Semantic Similarity Model (LQSSM) to deal with the problems that semantic information of summaries is incomplete and assessing the linguistic quality difficultly.
- The proposed linguistic quality evaluation network (LQEN) and semantic similarity evaluation network (SSEN) applied neural networks to evaluate summaries, which could reduce manual workload while improving the accuracy of evaluation.
- According to the characteristics of summaries in different aspects, this paper proposed an efficient method to construct datasets.

* Corresponding author: Bo Liu

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00113

Mainly focusing on similarity

II. RELATED WORK

A. Summary evaluation methods

In recent years, researchers have done a lot of works on summary evaluation methods. The early evaluation metrics are recall, precision and F-measure [4, 5]. Recall (R) and Precision (P) are count the number of sentences that appear in both peer and reference summaries, then divided by the number of sentences in reference summaries or peer summaries. F-score is a composite measure that combines precision and recall. Saggio et al. [6] introduced three automatic evaluation methods: cosine similarity, unit overlap and longest common subsequence, they are content-based similarity measures. Chin-Yew Lin [7] proposed Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which based on N-gram co-occurrence information, it evaluated summaries by counting the number of basic units that overlap between automatic summaries and reference summaries. ROUGE has become one of the general standards of text evaluation method, including ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, etc. Elghannam et al. [8] presented a new Keyphrase based Summary Evaluator (KpEval) to evaluate automatic summaries. The idea of KpEval is utilize three models to count the number of matches between the essential parts of system summaries and human summaries. Epitler et al. [9] focused on the measure of linguistic quality in summaries, combined different types of features such as modeling language, entity grid to evaluate five linguistic properties: referential clarity, focus, grammaticality, non-redundancy, and structure and coherence, but their study uninvolved the content of summaries.

B. BERT

We can consider the summary evaluation as a binary classification task in machine learning that use the model to classify the summaries into good and bad then get the corresponding score. The development and successful application of deep learning [10-12] have made the related technologies to complete the classification task get better results.

Bidirectional encoder representation from Transformers (BERT) is a large-scale pre-training model proposed by the Google team in 2018 [13]. It utilizes deep bidirectional Transformer encoder to mine the deep information of the language. BERT can be used for many natural language processing tasks due to it combines the two tasks of pre-training process. Firstly, drawing on the idea of cloze, masked language model task randomly masks some percentage of the input tokens, then predicts the original vocabulary of masked tokens by the encoder. Secondly, in order to understand the relationship between two sentences, next sentence prediction task was proposed. It means that the model receives pairs of sentences A and B as input, then the model predicts whether two sentences have a contextual relationship.

The application of BERT is convenient [14, 15], by fine-tuning the additional output layer it can create adaptive models for other tasks, such as classification, question answering, and so on. That only requires a small number of datasets. Therefore, based on the BERT model, this paper performed targeted processing on the classifier of downstream tasks to propose our model.

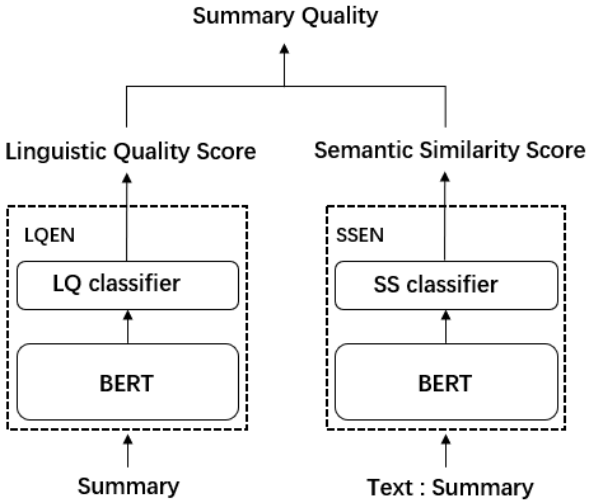

Figure 1. The architecture of LQSSM.

III. SUMMARY EVALUATION MODEL

This paper proposed LQSSM to evaluate the summary by combining linguistic quality and semantic similarity. As shown in Fig. 1, it was mainly composed of two networks, namely linguistic quality evaluation network and semantic similarity evaluation network, they have the same structure which includes BERT extractor and classifier. In the classifier part, we designed different networks to evaluate summaries from the perspective of grammar and semantics more accurately. Then, for measuring the summary as a whole, we merged the indicators of two evaluation network into summary quality (SQ) score which can evaluate the text summarization more intuitively.

A. Linguistic quality evaluation network

There are often have problems with grammatical errors or repeated words in generated summary [3]. Such a summary is unqualified and should be accurately identified in evaluation. The current solution is to start with the model that generates the summary and purpose is solving the problem completely. The summary evaluation method still follows the ROUGE metric [7], but it cannot directly reflect the impact of above problems.

In this paper, LQEN was introduced to address those problems, as shown in Fig. 2. We defined the task of evaluation network as a binary classification problem, and taken the summary to be evaluated as input of the network. Then extracted deep features of the summary through the pre-trained model BERT which was the extractor of the network. The extracted features C was mapped to high dimensions by the feed forward network and the activation function. Simultaneously, we used dropout [16] which could discard certain neurons with a probability to prevent overfitting. Finally, through dimensionality reduction by the non-linear activation function we got the output with two dimensions,

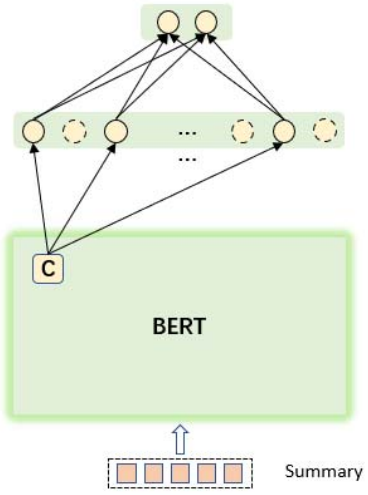

Figure 2. Linguistic quality evaluation network.

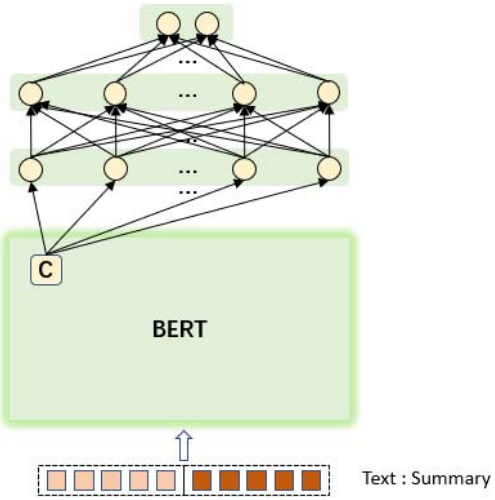

Figure 3. Semantic similarity evaluation network.

which can score the linguistic quality of the summary. The classifier of LQEN can be defined as:

$$r \sim Bernoulli(p), \tag{1}$$

$$y = sigmoid(r * \sigma(WX + b)). \tag{2}$$

Where r represents a binary vector 0 or 1 and obeys the Bernoulli distribution with probability p. σ is the activation function ReLU, which can introduce nonlinear factors. W is the weight vector that connecting the output of the pre-trained model and the input of the classification section, and use Xavier [17] method to initialize. X represents the input vector, b is the bias, we set it to 0.1. The sigmoid function is used as the activation function for obtaining the final score.

B. Semantic similarity evaluation network

The main purpose of summary evaluation is to determine whether the summary can accurately express the meaning of texts. While the central idea of texts can be expressed in multiple ways. For summaries with different expressions but the same semantics, the evaluation method should get similar results. However, the current summary evaluation method is mainly based on the word co-occurrence rate of the human answer and the generated summary. Faced with the above problems it will give distinct scores and this is not the outcome we expected.

In this paper, we proposed SSEN which was still fine-tuning the downstream networks based on the pre-trained model, as shown in Fig. 3. The difference is that discriminative learning of sentence pairs was added to the SSEN. Specifically speaking, we used the pairs of original text and summary as input, and modelled the two sentences, that could focus on extracting the semantic relationship between the original text and the summary. Then, the features extracted by BERT model were transferred to the two-layer fully connected network, which mapped the vector to the semantic space of the corpus. Finally, utilized the identical method to reduce dimension and obtained the prediction with two dimensions. The classifier of SSEN can be defined as:

$$y = sigmoid(\sigma(W_2\sigma(W_1X + b_1) + b_2)). \tag{3}$$

Where X represents the input vector, W_1 represents the parameter vector from the general semantic space to the similar semantic space, b_1 is the bias, W_2 represents the parameter vector for further conversion from the similar semantic space, and b_2 is also the bias. The initialization methods of W_1 and W_2 are same as W, and the expressions of σ and sigmoid are same as above.

C. Summary quality score

To have an overall evaluation of summaries, we integrated the linguistic quality score and the semantic similarity score to obtain a comprehensive score SQ. The method used in this paper is to multiply the score of them, so as to ensure that the high-score summary has a relatively high score in linguistic quality and semantic similarity. If one item fails, the SQ will fail, which is also in line with actual usage scenarios.

IV. Experiment

A. Datasets

The model training data was built on the basis of Large-scale Chinese Short Text Summarization (LCSTS) dataset, which is introduced by Hu et al. [18] of Harbin Institute of Technology. LCSTS dataset contains more than 2 million real Chinese short texts data and corresponding short summaries given by the writer. The author team also manually annotated 10666 text summaries. We selected a part of LCSTS dataset to construct our dataset in the next section, and finally obtains the linguistic quality datasets and semantic similarity dataset, as shown in Tables 1 and 2.

TABLE I. LINGUISTIC QUALITY DATASETS

	Training data	Test data	Validation data
Totals	2428	50	50
Positive samples	1205	24	21
Negative samples	1223	26	29

TABLE II. SEMANTIC SIMILARITY DATASETS

	Training data	Test data	Validation data
Totals	40000	1000	1000
Positive samples	20000	500	500
Negative samples	20000	500	500

B. *Datasets construction*

For the positive samples of linguistic quality dataset and semantic similarity dataset, the construction methods are identical. We directly select normal human summaries from LCSTS as positive samples. Since the data comes from a large-scale public dataset that its summaries were written by humans, the quality of summaries is reliable. The negative samples should reflect the true distribution of poor summaries, so we use different methods to obtain according to the characteristics of two aspects.

In the early preparations, we have been utilized several models of text summarization to generate summaries. Through experiments, we found that the Seq2Seq model which based on Recurrent Neural Network (RNN) [19] has the worst result in generating summaries, the linguistic quality of summaries hardly reaches the qualified standard. Therefore, the method for obtaining negative samples of linguistic quality dataset is known: use the Seq2Seq model based on RNN to infer some data which as the candidate set of negative samples, then human filter the data to get summaries that meet requirements. The samples of linguistic quality datasets are shown in Tables 3 and 4.

In generated summaries, there is a problem of incomplete semantic information. For example, the generation over with a half sentence, or the summaries only express partial information about texts. Considering above problems, we adopt two strategies to construct negative samples of the semantic similarity dataset. One is to cut summaries taken from LCSTS. We keep only 50% of the original summary length, and half retains the first 50%, the other retains the last 50%. The other is to randomly replace summaries to construct mismatched pairs of source text and summary. The samples of semantic similarity datasets are shown in Table 5.

TABLE III. LINGUISTIC QUALITY DATASETS SAMPLES

Positive samples	Negative samples
《星际穿越》里的黑洞和虫洞真的存在吗？	《星际穿越》：濒临地球毁灭
微博卖手机新浪要改行？	新浪微博将推出新浪微博专场预约？
投资圈的熟人经济：介绍人带来真金白银	“二八法则”是“二八法则”吗？

TABLE IV. TRANSLATION OF LINGUISTIC QUALITY DATASETS SAMPLES

Positive samples	Negative samples
Do the black holes and wormholes in *Interstellar* really exist?	*Interstellar*: on the verge of destruction of the earth
Weibo sells mobile phones, Sina wants to change business?	Sina Weibo will launch Sina Weibo special appointment?
Acquaintance economics in the investment circle: introducers bring real money	Is "Pareto principle" a "Pareto principle"?

TABLE V. SEMANTIC SIMILARITY DATASETS SAMPLES

Positive samples	Negative samples
特斯拉无钥匙驾驶过程可以更简单	特斯拉无钥匙驾驶
家电专业店营收增长净利润下滑	增长净利润下滑
亚洲资源企业：创造平衡的全球运营模式	平衡的全球运营模式

C. *Experiment setting*

The model training was fine-tuning based on the pre-trained model which is BERT-base with total of 12 layers, the hidden layer was 768 dimensions, and the self-attention heads was 12. In the linguistic quality evaluation network, we set the

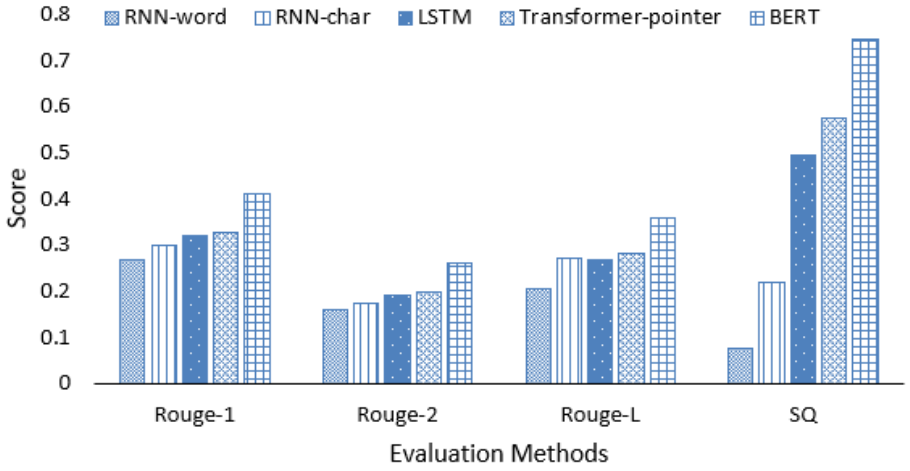

Figure 4. Compare the scores of summaries generated by different models on SQ and other evaluation methods.

cell numbers of feed forward network to 1024, and the value of dropout was set to 0.9. In the semantic similarity evaluation network, the cells of two-layer fully connected network were set to 1024 and 512 respectively. Moreover, when training model, we used a batch size of 16 and set the epoch to 2. The learning rate was set to 5e-5 and Adam [20] was used for optimization, the parameters were taken default values. All the experiments were carried out on a machine with GeForce GTX 1080, memory 16 GB, and Ubuntu 18.04.1 LTS operate system.

V. RESULT AND DISCUSSION

In order to validate the effectiveness of our method, we have compared five models to generate 726 summaries separately, including RNN-char, RNN-word [21], Transformer-pointer [22], Long Short-Term Memory (LSTM) [23] and BERT. Then, we measured our method from two aspects. The first, we scored these summaries using LQSSM, and compared with the scores of ROUGE-1, ROUGE-2 and ROUGE-L. The second, by analyzing the scores from different perspectives, our method further given a fine-grained evaluation to guide the optimization of summary models.

A. Comparison with different methods

It can be seen from Fig. 4, the score ranges of summaries on the four methods are [0.268, 0.410], [0.161, 0.262], [0.204, 0.360], and [0.077, 0.747]. Compared with the concentrated intervals on ROUGE-1, ROUGE-2, and ROUGE-L, the score intervals on SQ are more significantly dispersed. Additionally, for the summaries generated by Transformer-pointer and LSTM, the score differences of ROUGE metrics are 0.007, 0.005 and 0.015 respectively, while the difference of SQ is 0.081, which is larger. It is easy to find that in measuring the performance of models, the SQ and ROUGE metric are consistent, however compared with the three methods of ROUGE, SQ metric has an ability to distinguish the model more obviously.

B. Analysis of the scores

In order to clarify that SQ score can give corresponding assess and feedback on the ability of generated models from two perspectives, we made statistics on the linguistic quality scores and semantic similarity scores of summaries generated

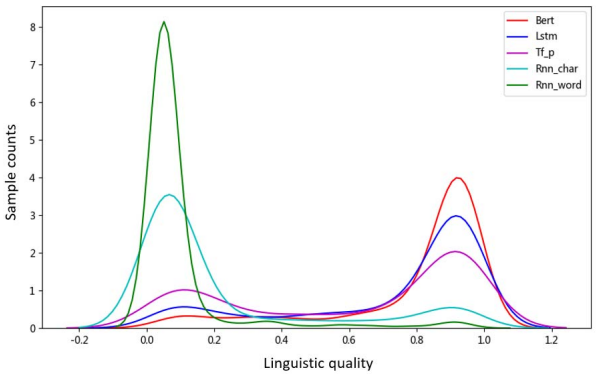

Figure 5. The distribution of summary linguistic quality scores.

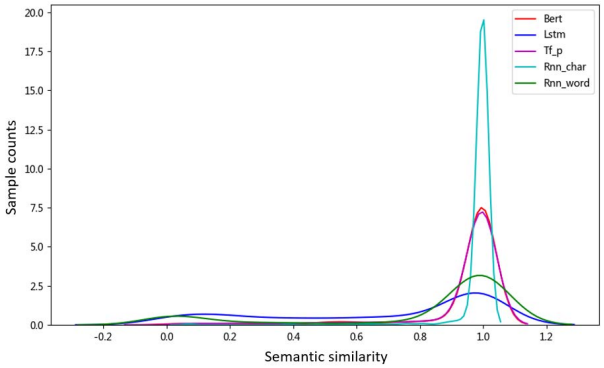

Figure 6. The distribution of summary semantic similarity scores.

by several models and produced frequency distribution graphs, which are shown in Figs. 5 and 6, respectively. The horizontal coordinate of the graph represents the range of scores which between 0 and 1, and the vertical coordinate represents the number of samples that get the corresponding score. The peak indicates that there are more samples distributed in this scoring area. Therefore, the more peaks in high-score partition, the better performance of the summary generation model in this respect, and the stronger generate ability.

From Figs. 5 and 6, we can see that BERT model performs well in both aspects, because the scores are mainly concentrated in the high-score range. For Transformer-pointer, the semantic performance is close to BERT, while the relatively poor performance on linguistic quality which has more low-score range. In contrast, LSTM has the generate ability second only to BERT on linguistic quality, but its semantic performance is not stable enough. Compared with them, two RNN models are not satisfactory. Their linguistic quality scores are concentrated in low-score areas, indicating that the linguistic quality of summaries is pretty bad. Simultaneously, RNN models performs well from a semantic perspective, RNN-char is even better than BERT. However, analyzing the specific samples, we found that the summaries generated by two RNN models are mainly the repetition of certain words, which causes a high score. This also shows that the evaluation cannot be based solely on semantics, linguistic quality also should be combined.

VI. CONCLUSIONS

In this paper, we proposed a summary evaluation method LQSSM which evaluated summaries combining the two dimensions of linguistic quality and semantic similarity. To train the model, we aimed at the characteristics of summaries, introduced an unsupervised learning combined with artificial methods to build datasets. The experimental results show that the method proposed in this paper can not only evaluate the similarity between summaries and original texts, but also automatically evaluate the linguistic quality of summaries. It has an ability to evaluate more comprehensive and detailed, that can provide appropriate feedback for summaries. Moreover, when scoring summaries there is no need to use human summaries for comparison, which effectively reduces manual workload. The future research can optimize the

semantic similarity network to reduce the semantic similarity score, make the semantic distinction more obviously.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61702021), and the grant of China Scholarship Council.

REFERENCES

[1] I. Mani and M. Maybury, Advances in Automatic Text Summarization, MIT Press, 1999.

[2] K. S. Jones and J. Galliers, Evaluating Natural Language Processing Systems: An Analysis and Review, Springer-Verlag Press, 1995.

[3] E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," Language Resources and Evaluation, vol. 52(1), 2018, pp. 101–148.

[4] D. Marcu, "The automatic construction of large-scale corpora for summarization research," Proc. ACM SIGIR Int. Conf. Information Retrieval, California, pp. 137–144, August 1999.

[5] H. Jing and K. McKeown, "The decomposition of Human-Written summary sentences," Proc. ACM SIGIR Int. Conf. Information Retrieval, California, pp. 129–136, August 1999.

[6] H. Saggion, D. Radev, and S. Teufel, "Meta-evaluation of summaries in a cross-lingual environment using content-based metrics'. COLING Int. Conf. Computational Linguistics, Stroudsburg, USA, pp. 1–7, 2002.

[7] C. Lin, "ROUGE: a package for automatic evaluation of summaries," In proceedings of the ACL 2004 Workshop on Text Summarization Branches Out, pp. 74–81, 2004.

[8] F. Elghannam and T. El-Shishtawy, "Keyphrase based evaluation of automatic text summarization," International Journal of Computer Applications, vol. 117(7), 2015, pp. 5-8.

[9] E. Pitler, A. Louis, and A. Nenkova, "Automatic evaluation of linguistic quality in multi-document summarization," Proc. ACL, Uppsala, Sweden, pp. 544–554, 2010.

[10] M. Peters, M. A. Neumann, and M. Iyyer, "'Deep contextualized word representations," arXiv preprint arXiv: 1802.05365, 2018.

[11] T. Young, D. Hazarika, and S. Poria, "'Recent trends in deep learning based natural language processing," IEEE Computational Intelligence Magazine, vol. 13(3), 2018, pp 55-75.

[12] J. Howard and S. Ruder, "Fine-tuned language models for text classification," arXiv preprint arXiv:1801.06146, 2018.

[13] J. Devlin, M. A. Chang, and K. Lee, "BERT: pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[14] C. Sun, X. Qiu, and Y. Xu, "How to fine-tune BERT for text classification?" arXiv preprint arXiv: 1905.05583, 2019.

[15] W. Yang, H. Zhang, and J. Lin, "Simple applications of BERT for Ad Hoc document retrieval," arXiv preprint arXiv: 1903.10972, 2019.

[16] N. Srivastava, G. Srivastava, and A. Krizhevsky, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research (JMLR), vol. 15(56), 2014, pp 1929−1958.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proc. PMLR Int. Conf. Artificial Intelligence and Statistics, Sardinia, Italy, pp 249-256, 2010.

[18] B. Hu, Q. Chen, and F. Zhu, "LCSTS: a large scale chinese short text summarization dataset," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, pp. 1967–1972, 2015.

[19] K. Cho, B. Van Merriënboer, and C. Gulcehre, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[20] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," International Conference on Learning Representations (ICLR), San Diego, USA, May 2015.

[21] Karpathy and Andrej, "The unreasonable effectiveness of recurrent neural networks" Andrej Karpathy blog, pp. 21–23, 2015.

[22] S. Abigail, and J. Peter, "Get to the point: summarization with pointer-generator networks," arXiv preprint arXiv: 1704.04368, 2017.

[23] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9(8), 1997, pp. 1735-1780.