Journal of Computer and Communications, 2020, 8, 1-16
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Intrinsic and Extrinsic Automatic Evaluation Strategies for Paraphrase Generation Systems

Tulu Tilahun Hailu[1], Junqing Yu[1,2*], Tessfu Geteye Fantaye[1]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[2]Center of Network & Computation, Huazhong University of Science and Technology, Wuhan, China
Email: tutilacs@yahoo.com, *yjqing@hust.edu.cn, tessfug@hust.edu.cn

How to cite this paper: Hailu, T.T., Yu, J.Q. and Fantaye, T.G. (2020) Intrinsic and Extrinsic Automatic Evaluation Strategies for Paraphrase Generation Systems. *Journal of Computer and Communications*, **8**, 1-16.
https://doi.org/10.4236/jcc.2020.82001

Received: January 12, 2020
Accepted: February 8, 2020
Published: February 11, 2020

Abstract

Paraphrase is an expression of a text with alternative words and orders to achieve a better clarity. Paraphrases have been found vital for augmenting training dataset, which aid to enhance performance of machine learning models that intended for various natural language processing (NLP) tasks. Thus, recently, automatic paraphrase generation has received increasing attention. However, evaluating quality of generated paraphrases is technically challenging. In the literature, the importance of generated paraphrases is tended to be determined by their impact on the performance of other NLP tasks. This kind of evaluation is referred as extrinsic evaluation, which requires high computational resources to train and test the models. So far, very little attention has been paid to the role of intrinsic evaluation in which quality of generated paraphrase is judged against predefined ground truth (reference paraphrases). In fact, it is also very challenging to find ideal and complete reference paraphrases. Therefore, in this study, we propose semantic or meaning oriented automatic evaluation metric that helps to evaluate quality of generated paraphrases against the original text, which is an intrinsic evaluation approach. Further, we evaluate quality of the paraphrases by assessing their impact on other NLP tasks, which is an extrinsic evaluation method. The goal is to explore the relationship between intrinsic and extrinsic evaluation methods. To ensure the effectiveness of proposed evaluation methods, extensive experiments are done on different publicly available datasets. The experimental results demonstrate that our proposed intrinsic and extrinsic evaluation strategies are promising. The results further reveal that there is a significant correlation between intrinsic and extrinsic evaluation approaches.

Keywords

Paraphrase, Paraphrase Generation, Natural Language Processing, Intrinsic, Extrinsic, Automatic Evaluation, Word Embedding, Sentiment Analysis

*Corresponding author.

1. Introduction

Paraphrase is an expression that delivers the same information of the original text by using different words and order [1]. It has been proved that paraphrases play a vital role for augmenting and bringing diversity to existing training dataset, which significantly improves performance of machine learning models that intended for NLP tasks. Machine learning models inherently require large training dataset. If the training sample is big enough, the models can easily capture all the discrepancies and learn relevant patterns. Failing to find enough dataset obscures application of machine learning techniques to natural language processing tasks, especially for low-resource and morphologically rich languages [2]. Hiring people to collect large dataset is not practical. Alternatively, large volume of synthetic data can be used to train complex machine learning models. Rubin [3] discussed the validity of synthetic data with the intention of preserving confidentiality of authentic data. Barse *et al.* [4] also used generated artificial data to train fraud detection system.

Similarly, automatically generated paraphrases have been used to improve various NLP models, for example, question-answering [5] [6]; information extraction and retrieval [7] [8]; relation extraction [9]; text summarization [10] [11]; machine translation [12] [13] [14]; automatic generation of reference translation [15] [16], etc. As a result, recently, in the field of NLP, automatic paraphrase generation received increasing attention. In line with this, quality of generated paraphrases could be determined based on the impact of generated paraphrases on the performance of target NLP task.

According to Alexander *et al.* [17], perhaps automatic and manual evaluations are the most basic dichotomy for evaluating NLP systems. To assess performance of NLP systems, recruiting human experts is the most straightforward. However, it has two major limitations: first, humans often generate inconsistent results and the process is very slow. Secondly, manual evaluation is a time-consuming task and laborious. In other words, it is a costly evaluation approach. Thus, automatic evaluation is relatively more vital to evaluate and rank several NLP systems.

Evaluation of NLP systems can also be classified into intrinsic and extrinsic methods [17], which can be performed either automatically or manually. In an intrinsic evaluation, quality of NLP systems outputs is evaluated against pre-determined ground truth (reference text) whereas an extrinsic evaluation is aimed at evaluating systems outputs based on their impact on the performance of other NLP systems. For instance, in intrinsic evaluation of paraphrase generation system, we would ask the following questions: Does the generated paraphrase convey meaning of the original text? On the other hand, extrinsic evaluation of generated paraphrases deals with their impact on the performance of other NLP systems. In this context, we might ask: Do the incorporated paraphrases significantly improve performance of question-answering model? Can the generated paraphrases be used as a surrogate of original text to train text classification models? If so, it can be concluded that, extrinsically, the considered paraphrases are useful.

If the goal is to evaluate quality of generated paraphrases or performance of paraphrase generation systems, then it is usually easier to use automatic evaluation metric for intrinsic evaluation [17] [18]. However, automatic intrinsic evaluation of paraphrase generation systems is also becoming harder for two main reasons: first, it is difficult to find an acceptable ground truth (reference paraphrases). It is very challenging for human to produce complete and ideal reference sentences or phrases [19]. In other words, as previously proposed automatic evaluation metrics rely on lexical matching, if the terms or fragments in the generated paraphrases do not exist in the human generated reference paraphrase, despite their goodness, those systems will be penalized which is unfair.

Secondly, because of various characteristics of paraphrases, the generated paraphrases that have high metric score with the corresponding reference paraphrases might be not genuine. For example, if a paraphrase generation system generates an exact copy of the original sentence or phrase, based on the fundamental definition of paraphrase, the generated text is actually not a paraphrase rather copy of the original text. Thus, the system that generates an exact copy of the original text deserves a low metric score. However, researchers have not treated this situation in much detail.

Thus, in this study, we propose an appropriate automatic evaluation metric for paraphrase generation systems, which is intrinsic evaluation approach. In the proposed metric, we incorporate a way of discouraging copying of words or phrases from the original text and encouraging using of alternative words in the generated paraphrases. Further, we use generated paraphrases to train different NLP models and assess their impact, which is an extrinsic evaluation method. The aim is to explore relationship between intrinsic and extrinsic automatic evaluation of paraphrase generation systems. Our main contribution can be summarized as follows:

- Based on lesson learned from study in [14], we generate paraphrases and prepare them for analyzing intrinsic and extrinsic evaluation strategies and assess their relationships.
- For intrinsic evaluation, we propose a suitable automatic evaluation metric for paraphrase generation system.
- For extrinsic evaluation, we make use of generated paraphrases and train various sentiment classification models and assess their impact on the performance of these models.
- We conduct extensive experiments on different publicly available dataset configurations. The experimental results show that the proposed intrinsic and extrinsic evaluation strategies are promising. The results also demonstrate that there is a significant correlation between intrinsic and extrinsic evaluation of paraphrase generation systems.

The remainder of this paper is organized as follows: In Section 2, first we review related works in terms of paraphrase generation approaches and paraphrase evaluation methods. In Section 3, we describe the proposed intrinsic and

extrinsic evaluation strategies and their implementation details. Experimental results and discussion are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Work

In this section, we review the related work from the following two perspectives: paraphrase generation and evaluating quality of generated paraphrases.

2.1. Paraphrase Generation

Paraphrase generation (PG) is a process of presenting and conveying information of original sentence/phrase in alternative words and order [17] [20]. Paraphrase generation approaches may be divided into two main categories. 1) rule-based; and 2) machine learning based approaches.

Rule-based: In rule-based PG, rules are created manually to transform original text into semantically equivalent text or paraphrases [21] [22]. Most of manually crafted rules were intended to use linguistic resources such as dictionaries, WordNets or thesaurus for replacing words in the original text with their synonyms [16] [23] [24]. Words in the original text can also be replaced with their antonyms along with appropriate negation words. Further, linguistic structure has been leveraged for generating paraphrases. For instance, changing active voice into passive; adding or deleting function words; co-reference substitution; and changing part-of-speech, just to mention some. In this regard, study in [1] identified 25 paraphrase generation techniques. Moreover, based on lexico-grammatical resources, Raymond *et al.* [25] discussed ways of generating paraphrases from predicate/argument. Study in [26] extended the usage of linguistic resources at lexical and syntactic levels to document structure and layout for paraphrase generation.

Machine learning based: In machine learning based PG, rules that help to generate paraphrases are created automatically from the data [19] [27] [28]. Scheme in statistical machine translation (SMT) has been adopted for paraphrases generation [29] [30]. The main difference between them is that in SMT source and target language texts are from different languages (bilingual) whereas in paraphrase generation both source and target side texts are from the same language (monolingual). Further, various advanced machine learning approaches have also been adopted for paraphrase generation. For instance, deep neural network [31], deep reinforcement [32], generative adversarial network (GAN) [33], and combining deep generative variation auto encoder (VAE) with sequence-to-sequence long short-term memory (LSTM) models [34]. However, to train the deep learning neural network models, scarcity of training dataset is noticeably a challenging issue. It should be noted that generative models are capable to generate paraphrases that totally different from the original text in surface form but semantically similar. In this regard, however, existing automatic evaluation metrics are not appropriate to judge quality of generated paraphrases be-

cause the most commonly used automatic evaluation metrics rely on lexical matching. Thus, to fairly assess performance of generative models for paraphrase generation, semantic oriented automatic evaluation metrics would be vital.

2.2. Evaluating Paraphrase Generation Models/Systems

Evaluation of NLP models or systems is a more general term. Philip and Jimmy [18] discussed various categories of NLP systems evaluation approaches. According to them, there are several evaluation dichotomies: automatic versus manual; formative versus summative; intrinsic versus extrinsic; and component versus end-to-end. Further, there is a kind of evaluation called adequacy and fluency. These evaluation approaches can also be applied to specific NLP task including paraphrase generation.

According to Philip and Jimmy, recruiting humans and ask them to assess quality of NLP systems output on the basis of pre-determined criteria is the most straightforward. However, it has two main limitations: firstly, human judgments are notoriously inconsistent. In other words, there is a low inter-annotators agreement. Secondly, manual evaluation approach is a time-consuming and laborious task. Thus, automatic evaluation approach is increasingly receiving more attention.

Evaluation can also be considered as part of development life cycle of NLP system because all developers need to evaluate the system under study at the modular level, which is referred as formative evaluation. On the other hand, evaluation of the systems at the final stage, after integrating the whole modules but before deploying to the real working environment, is referred as summative evaluation. Formative and summative evaluations are very similar with that of component and end-to-end evaluations respectively. Component based evaluation is assessing quality of the system at individual component rather than evaluating the system as a whole at once. For instance, in order to evaluate parsers that developed based on part-of-speech tagger and other components, the component based evaluation is intended to evaluate the parser based on the part-of-speech tagger first and also based on other components, individually. In contrast, in the end-to-end evaluation, quality of the whole system is evaluated at once.

The most commonly invoked NLP systems evaluation approaches are intrinsic and extrinsic. In intrinsic evaluation, system output is evaluated against the pre-determined ground truth (reference text) whereas in extrinsic evaluation quality of system output is assessed based on its impact on the performance of other NLP systems [35] [36]. Intrinsic evaluation can be further divided into two main categories: adequacy and fluency. Adequacy is aimed to judge how much of the meaning expressed in the ground truth is preserved in the system output whereas fluency dealt with the intuitively acceptable linguistic structure or grammatical correctness of systems outputs. In extrinsic based evaluation methods (aka task-based evaluation), quality of generated texts (paraphrases) can-

not be determined until the generated paraphrases are used for training and testing the models. Accordingly, one can judge quality of various generated paraphrases or performance of paraphrase generation systems after observing the improvement or validity of generated paraphrases to be a surrogate for the original texts.

If the basic intention is to assess quality of the paraphrases, task-based evaluation might not be a good choice. Because, training and testing of machine learning models requires high computational power and large dataset, which is not available in most cases. Thus, intrinsic method or automatic evaluation metric for paraphrase generation systems would be more preferable. However, as far as we know, only three automatic evaluation metrics have been proposed to assess performance of paraphrase generation systems: ParaMetric [37], PEM (Paraphrase Evaluation Metric) [38], and PINC (Paraphrase In N-gram Changes) with BLEU (Bilingual Evaluation Understudy) [39].

ParaMetric automatic evaluation metric is similar with the Pyramid method [40] that proposed to evaluate performance of text summarization systems. Both are intended to complement automatic evaluation by employing human annotators for introducing varieties to the reference paraphrases. Even though it is a good attempt, it is very challenging to prepare complete and ideal reference paraphrases. As an alternative, to establish semantic equivalence, a trainable metric called PEM was proposed to use a second language as a pivot. Although PEM was shown to correlate well with human judgments, the requirement of large parallel text for training is as difficult as generating paraphrases. On the other hand, PINC was proposed to assess dissimilarity between paraphrase and source text. For dissimilarity measurements, PINC score is computed in an opposite to the BLEU. Further, the authors also used BLEU for measuring adequacy and fluency of generated paraphrases against the source text. Accordingly, the authors used BLEU and PINC as a two dimensional scoring metric, which is logical. However, the measurement is solely relying on lexical matching by ignoring semantic similarities, which can be a fundamental defect.

In the literature, the BLEU metric [41], which was originally proposed for assessing quality of machine translation systems has been used for evaluating quality of generated paraphrases. BLEU metric relies on n-gram lexical matching, which has several flaws. It fails to evaluate generated text based on their meaning and the considered geometric mean of n-units in BLEU ends up with unfair score. This limitation is severe when it comes to the evaluation of paraphrase generation systems. For instance, evaluation of generated paraphrases against reference paraphrases by using lexical matching based metrics might not aligned with the fundamental definition of paraphrase. The basic definition of good paraphrase is that it should be lexically dissimilar with the source text while preserving its meaning. In other words, it seems better to evaluate the generated paraphrases against the original text. What if paraphrase generation model/system generate an exact copy of original text. Obviously, the generated text cannot be

considered as a paraphrase it is a copy of the original text. Thus, it is very important to consider this criterion in the development of automatic evaluation metric for PG, like study in [39].

In some studies, researchers used METEOR metric [42] rather than BLEU, which was also originally proposed to evaluate MT systems. Stemming and synonym matching techniques have been incorporated in METEOR, which helps to alleviate problems of exact matching in BLEU. However, METEOR requires linguistic resources such as parsers and WordNet, which are very difficult to find fully fledged linguistic resources for most languages. It would be better to leverage other linguistic resources like publicly available pre-trained word embedding models that proved to capture syntactic, semantic, and morphological similarity of the words [43] [44]. Noticeably, word embedding models are increasing published for several languages [45] and even relatively easier to develop from scratch.

Therefore, in this study, we propose a suitable semantic oriented automatic evaluation metric for paraphrase generation systems. Further, we also evaluate quality of generated paraphrases based on various sentiment classification models. Consequently, we conduct a correlation analysis between the automatic evaluation metric scores and task-based evaluation results. We describe procedure of doing this in the next section.

3. Method

In the literature, intrinsically, the BLEU metric has been commonly used to evaluate quality of paraphrases. Extrinsically, quality of generated paraphrases has also been assessed based on their impact on various NLP tasks such as question-answering, information retrieval, text summarization, machine translation etc. Although intrinsic evaluation is relatively more practical, the commonly used automatic evaluation metrics are not suitable to evaluate performance of paraphrase generation systems. Thus, we propose simple but yet more appropriate and effective automatic evaluation metric for paraphrase generation system and analyze relationship it has with the task-based automatic evaluation methods.

To achieve the stated goals, we performed the following three activities:

1) We adopted existing paraphrase generation method in [14] for generating paraphrases.

2) Based on the various characteristics of the paraphrases, we proposed a suitable automatic evaluation metric for evaluating quality of paraphrases or for evaluating performance of paraphrase generation systems.

3) For comparison, we identified some sentiment classification models that help use to assess quality of generated paraphrases. Consequently, we calculated the correlation between intrinsic and extrinsic evaluation results. The overall workflow is depicted in **Figure 1** and we describe each process in the next consecutive sections.

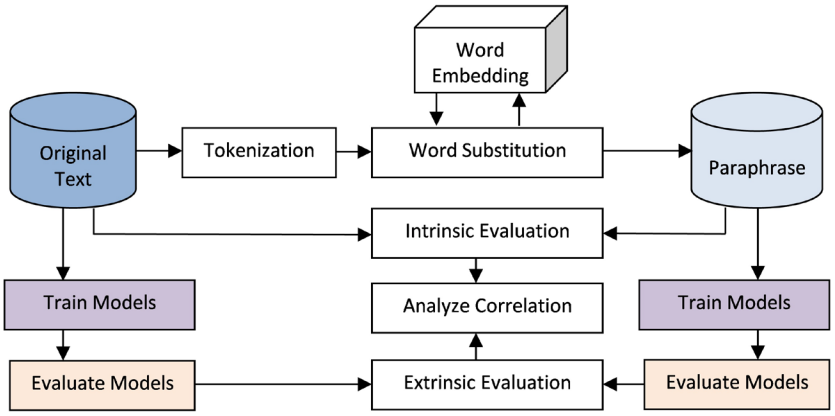

Figure 1. The general workflow architecture. It should be noted that the hyper-parameter settings of the models trained on the original dataset are the same with the models trained on the paraphrases dataset.

3.1. Dataset

In this study, we used datasets from two sources: first, for evaluating proposed automatic evaluation metric, we used 26 k and 13 k annotated phrase-paraphrase pairs that randomly drawn from Paraphrase Database (PPDB) and Wikipedia respectively that made publicly available by Pavlick *et al.* [46]. Secondly, for evaluating quality of generated paraphrases, we used large movie reviews dataset[1] that shared by Maas *et al.* [47]. Accordingly, we randomly draw 5000 reviews from 50,000 large movie review dataset along with the class label annotated by human. Then, we divided this dataset into three parts: 2500 reviews for training; 2000 reviews for validation; and the remaining 1000 reviews for testing.

3.2. Paraphrase Generation Methods

Based on the text augmentation techniques proposed in [14], we generated three groups of paraphrases, which can be referred as paraphrase generation system 1, 2 and 3 (PG1, PG2, and PG3). The first system (PG1) is intended to generate paraphrases by replacing words of the original text with the first most similar word in the vector space whereas the second and the third systems are intended to generate paraphrases by replacing the words in the original text with the second and third most similar words in the vector space respectively. For this purpose, we used publicly available pre-trained word embedding model that trained on Word2Vec [48]. To determine the most similar word of the original word in a vector space, we make use of python library called gensim.

3.3. Automatic Evaluation Metric for Paraphrase Generation

In order to evaluate quality of generated paraphrases, intrinsically, we develop a word embedding based evaluation metric for paraphrase generation (WEEM4PG for short). In WEEM4PG, we opt to compare the generated paraphrases directly with the original text. By this, we address lack of human generated reference pa-

[1]https://ai.stanford.edu/~amaas/data/sentiment/

raphrases. Further, comparing quality of generated paraphrases against the original text is better than comparing against the reference text because it helps to minimize some discrepancies. For instance, paraphrase is a way of expressing the original text with alternative words but deliver the same meaning. To confirm this, it is better to use the original text rather than using reference paraphrases. Thus, in WEEM4PG, we set criteria in which we discourage the overlapping words and encourage substitution of the words with the alternative words, see Equation (1).

$$\text{WEEM4PG} = \left[\frac{\sqrt{N(O \cap P)}}{N(O \cap P)}\right]\left[\frac{\sum_{i=1}^{n} w_i}{\text{length}(P)}\right] \quad (1)$$

where $N(.)$ represents the number of overlapping words between paraphrase (P) and original text (O). The w_i is the cosine similarity value between a word at position "i" in the generated paraphrase and the nearest word among the words in the original text in the vector space.

The first part of Equation (1) is square root of number of overlapping words divided by the total number of overlapping words between the original and the paraphrase texts, which is incorporated to discourage severe overlapping words whereas the second part helps to encourage semantic similarities.

3.4. Identifying Models for Extrinsic Evaluation

In order to determine quality of generated paraphrases, extrinsically, we first train sentiment classification models on the original movie reviews and predict on the reserved test dataset. The prediction results were saved to be used later for comparison. Similarly, we train the models on the paraphrases that generated from the original movie reviews and predict on the same test dataset. Then, we compare prediction results of sentiment classification models that trained on the original and paraphrase reviews. Accordingly, we can judge the impact of generated paraphrases on these models or to determine whether the generated paraphrases can be a valid surrogate for the original reviews, which in turn helps to determine quality of generated paraphrases. For this purpose, we identified four types of sentiment classification models:

1) Simple Multi-Layer Perceptron Neural Network (SMPNN);
2) Multichannel Convolution Neural Network (MCNN);
3) Gated Recurrent Unit based Recurrent Neural Network (GRU-RNN) and;
4) Long Short-Term Memory based Recurrent Neural Network (LSTM-RNN).

All models were developed based on Keras python library. The network configurations of these models are similar on some hyper-parameter settings. For instance, in all networks, the input layer (embedding layer) is defined with the input and output dimension of 5000 and 300 respectively, and input length of 500. Similarly, loss of all models is computed based on the binary cross entropy. Adam optimizer and accuracy metric are used in all cases. Further, the number of epochs to train all models is 30. Batch size of MCNN is 16 whereas 128 for the

other three models. These hyper-parameter values were determined by trial-and-error process or tuning and we used the default values for the other hyper-parameters.

4. Results and Discussion

4.1. Correlation Results of Automatic Evaluation Metrics

To analyze correlation between automatic evaluation metrics and human judgments, we used human judgments of PPDB 2.0 that made publicly available by [46] and report the results in Table 1. The table compares correlation coefficients of our proposed automatic evaluation metric (WEEM4PG) and other automatic evaluation metrics. Accordingly, WEEM4PG is in the middle position among the 5 competing metrics. Supervised scoring model in PPDB 2.0 and cosine similarity that based on word embeddings of rare words are better than WEEM4PG. A possible explanation for this might be that scoring techniques of these two metrics were adjusted according to the characteristics of phrase-paraphrase pairs in PPDB. For instance, in the supervised learning, phrases reserved for testing might appear in the training set. Similarly, in the cosine similarity based metrics, the considered rare words are defined in the context of PPDB.

Apart from WEEM4PG, for calculating scores, all automatic evaluation metrics reported in Table 1 are limited to a phrase level. In other words, WEEM4PG is intended to evaluate paraphrases against the original text beyond the phrases. Further, unlike other metrics, WEEM4PG is proposed to evaluate paraphrases of all kind including paraphrases in PPDB.

4.2. Prediction Accuracy Results of Sentiment Classification Models

We assessed the impact of generated paraphrases by evaluating performance of sentiment classification models that trained on four dataset configurations: original text, paraphrases generated based on PG1, paraphrases generated based on PG2, and paraphrases generated based on PG3. We used validation dataset to fine-tune the models hyper-parameters whereas testing dataset to evaluate performance of the final models.

Table 1. Spearman's (ρ) correlation of automatic evaluation metrics for paraphrase generation on two datasets: sample from PPDB 2.0 and Wikipedia. The best score is highlighted bold. All correlation coefficients are significant at $P \leq 0.05$.

Source of dataset	PPDB	Wiki
#Phrase-Paraphrase pairs	26,455	13,954
P(paraphrase/phrase [49]	0.414	-
Heuristic scoring in PPDB 1.0 [50]	0.407	-
Cosine similarity based on word embeddings of rare words [46]	0.463	
Supervised scoring model in PPDB 2.0 [46]	**0.713**	-
WEEM4PG (Ours)	0.435	0.382

Table 2 compares prediction accuracy results of four different sentiment classification models. As can be seen from the table, MCNN performs best on all training datasets when compared to other counterpart models. Performance of all models except GRU-RNN that trained on the original reviews is best when compared to the same network configuration trained on the generated paraphrases, which is logical and expected. However, GRU-RNN that trained on the paraphrases of PG2 outperforms same model that trained on other datasets including the original dataset, which is somewhat counterintuitive and also difficult to justify.

Closer inspection of results in Table 2 shows the differences of prediction accuracy of the models trained on the original text are very small when compared to the prediction accuracy of the models trained on the generated paraphrases. Thus, it can be concluded that the generated paraphrases are valid to be a surrogate for the original reviews.

4.3. Correlation between Intrinsic and Extrinsic Evaluations

Table 3 presents the correlation between intrinsic and extrinsic evaluation results. In other words, this table illustrates association between WEEM4PG scores and prediction results of sentiment classification models. As can be seen from the table, Pearson's and Spearman's correlation between WEEM4PG scores are significantly correlated with the prediction accuracy results of MCNN on the

Table 2. Prediction accuracy of four sentiment classification models that trained on four dataset configurations. Testing dataset is common in all cases. The best score is highlighted in bold.

Training Set	SMPNN	MCNN	GRU-RNN	LSTM-RNN
Original review (baseline)	73.08	83.00	51.44	49.92
Paraphrase PG1	65.04	68.28	51.16	49.24
Paraphrase PG2	**65.76**	74.04	**66.00**	48.88
Paraphrase PG3	64.36	**74.68**	51.28	**49.36**

Table 3. Pearson's (r) and Spearman's (ρ) correlation between WEEM4PG scores and prediction accuracy results of four sentiment classification models. All models are trained on the original text and the trained models are used to predict paraphrases generated by PG1, PG2, and PG3 from test dataset. The single and double superscript stars indicate the correlation is significant at the $P = 0.05$ and $P = 0.01$ respectively. Best results are highlighted in bold.

Models	WEEM4PG, PG1		WEEM4PG, PG2		WEEM4PG, PG3	
	r	ρ	r	ρ	r	ρ
SMPNN	−0.007	−0.009	0.163*	0.121	−0.014	−0.039
MCNN	**0.206****	**0.301****	**0.286****	**0.271****	**0.061**	**0.063**
GRU-RNN	0.201**	0.155**	−0.078	−0.088	−0.066	−0.085
LSTM-RNN	0.033	0.079	0.076	0.010	−0.1	−0.154

paraphrases generated by PG1 and PG2. Further, although the correlation coefficient values are very low, WEEM4PG has also a significant correlation with SMPNN on paraphrases generated by PG2 and with GRU-RNN on paraphrases generated by PG1. However, no significant correlation was found between WEEM4PG and all sentiment classification models on paraphrases generated by PG3. Moreover, no evidence was found regarding correlation between WEEM4PG and LSTM-RNN on all test datasets. Thus, it can be concluded that intrinsic evaluation method can agree with at least some extrinsic evolution methods.

5. Conclusion

In this study, the aim was to propose intrinsic and extrinsic automatic evaluation strategies and to explore association between them. The obtained experimental results show that the proposed word embedding based automatic evaluation metric (WEEM4PG) is promising. The results also revealed that WEEM4PG has a significant correlation with accuracy prediction results of some sentiment classification models. Taken together, the results suggest that WEEM4PG is more practical for three main reasons: first, it intended to evaluate quality of generated paraphrases against the original text, which helps to address lack of human generated reference paraphrases. Second, relatively, it is simple to use and cheaper when compared to extrinsic evaluation methods. Third, in the evaluation process, WEEM4PG is designed to discourage lexical overlapping while encouraging semantic similarities between the original texts and generated paraphrases, which preserve the fundamental definition of paraphrase. However, as this study is tended to evaluate how much meaning of the original text is preserved in the generated paraphrases (adequacy), unfortunately, assessment of grammatical correctness or fluency of generated paraphrases was ignored. Thus, further work is needed to evaluate paraphrases on the basis of adequacy and fluency.

Acknowledgements

We gratefully acknowledge the granted financial support from the National Natural Science Foundation of China (No. 61572211, 61173114, 61202300).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bhagat, R. and Hovy, E. (2013) What Is a Paraphrase? *Computational Linguistics*, **39**, 463-472.

[2] Antony, P., Raj, H.B., Sahana, B., Alvares, D.S. and Raj, A. (2012) Morphological Analyzer and Generator for Tulu Language: A Novel Approach. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, Chennai, India, 3-5 August 2012, 828-834. https://doi.org/10.1145/2345396.2345531

[3] Rubin, D.B. (1993) Statistical Disclosure Limitation. *Journal of official Statistics*, **9**, 461-468.

[4] Barse, E.L., Kvarnstrom, H. and Jonsson, E. (2003) Synthesizing Test Data for Fraud Detection Systems. 19*th Annual Computer Security Applications Conference*, Las Vegas, NV, 8-12 December 2003, 384-394. https://doi.org/10.1109/CSAC.2003.1254343

[5] Duboue, P. and Chu-Carroll, J. (2006) Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, June 2006, 33-36. https://doi.org/10.3115/1614049.1614058

[6] Yin, P., Duan, N., Kao, B., Bao, J. and Zhou, M. (2015) Answering Questions with Complex Semantic Constraints on Open Knowledge Bases. *Proceedings of the* 24*th ACM International on Conference on Information and Knowledge Management*, Melbourne, 19-23 October 2015, 1301-1310. https://doi.org/10.1145/2806416.2806542

[7] Apresjan, J.D., Boguslavsky, I.M., Iomdin, L.L., Cinman, L.L. and Timoshenko, S.P. (2009) Semantic Paraphrasing for Information Retrieval and Extraction. In: Andreasen, T., Yager, R.R., Bulskov, H., Christiansen, H. and Larsen, H.L., Eds., *Flexible Query Answering Systems. FQAS* 2009. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 512-523. https://doi.org/10.1007/978-3-642-04957-6_44

[8] Zukerman, I. and Raskutti, B. (2002) Lexical Query Paraphrasing for Document Retrieval. *Proceedings of the* 19*th international conference on Computational linguistics*, **1**, 1-7. https://doi.org/10.3115/1072228.1072389

[9] Romano, L., Kouylekov, M., Szpektor, I., Dagan, I. and Lavelli, A. (2006) Investigating a Generic Paraphrase-Based Approach for Relation Extraction. 11*th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006.

[10] Madnani, N., Zajic, D., Dorr, B., Ayan, N.F. and Lin, J. (2007) Multiple Alternative Sentence Compressions for Automatic Text Summarization. *Proceedings of DUC*, Rochester, NY, April 2007, 1-8.

[11] Barzilay, R. and Mckeown, K.R. (2003) Information Fusion for Multidocument Summarization: Paraphrasing and Generation.

[12] Callison-Burch, C., Koehn, P. and Osborne, M. (2006) Improved Statistical Machine Translation Using Paraphrases. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 17-24. https://doi.org/10.3115/1220835.1220838

[13] Madnani, N., Ayan, N.F., Resnik, P. and Dorr, B.J. (2007) Using Paraphrases for Parameter Tuning in Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, 120-127. https://doi.org/10.3115/1626355.1626371

[14] Hailu, T.T., Yu, J. and Fantaye, T.G. (2019) Pre-Trained Word Embedding Based Parallel Text Augmentation Technique for Low-Resource NMT in Favor of Morphologically Rich Languages. *Proceedings of the* 3*rd International Conference on Computer Science and Application Engineering*, Sanya, 22-24 October 2019, 1-5. https://doi.org/10.1145/3331453.3361309

[15] Zhou, L., Lin, C.-Y. and Hovy, E. (2006) Re-Evaluating Machine Translation Results with Paraphrase Support. *Proceedings of the* 2006 *Conference on Empirical Methods in Natural Language Processing*, Sydney, 22-23 July 2006, 77-84.

https://doi.org/10.3115/1610075.1610087

[16] Kauchak, D. and Barzilay, R. (2006) Paraphrasing for Automatic Evaluation. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 455-462. https://doi.org/10.3115/1220835.1220893

[17] Clark, A., Fox, C. and Lappin, S. (2013) The Handbook of Computational Linguistics and Natural Language Processing. John Wiley & Sons, New York.

[18] Resnik, P. and Lin, J. (2010) Evaluation of NLP Systems. In: *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley, New York, 57.

[19] Lin, D. and Pantel, P. (2001) Discovery of Inference Rules for Question-Answering. *Natural Language Engineering*, **7**, 343-360.

[20] Zhao, S., Lan, X., Liu, T. and Li, S. (2009) Application-Driven Statistical Paraphrase Generation. *Proceedings of the Joint Conference of the* 47*th Annual Meeting of the ACL and the* 4*th International Joint Conference on Natural Language Processing of the AFNLP*, **2**, 834-842. https://doi.org/10.3115/1690219.1690263

[21] McKeown, K.R. (1980) Paraphrasing Using Given and New Information in a Question-Answer System. Technical Reports (CIS). https://doi.org/10.3115/982163.982182

[22] Zong, C., Zhang, Y., Yamamoto, K., Sakamoto, M. and Shirai, S. (2001) Approach to Spoken Chinese Paraphrasing Based on Feature Extraction. *NLPRS*, 551-556.

[23] Bolshakov, I.A. and Gelbukh, A. (2004) Synonymous Paraphrasing Using WordNet and Internet. In: Meziane, F. and Métais, E., Eds., *Natural Language Processing and Information Systems. NLDB* 2004. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 312-323. https://doi.org/10.1007/978-3-540-27779-8_27

[24] Mueller, J. and Thyagarajan, A. (2016) Siamese Recurrent Architectures for Learning Sentence Similarity. 13*th AAAI Conference on Artificial Intelligence*, Phoenix Convention Center, Phoenix, AZ, 12-17 February 2016.

[25] Kozlowski, R., McCoy, K.F. and Vijay-Shanker, K. (2003) Generation of Single-Sentence Paraphrases from Predicate/Argument Structure Using Lexico-Grammatical Resources. *Proceedings of the Second International Workshop on Paraphrasing*, **16**, 1-8. https://doi.org/10.3115/1118984.1118985

[26] Power, R. and Scott, D. (2005) Automatic Generation of Large-Scale Paraphrases.

[27] Barzilay, R. and Lee, L. (2003) Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. *Proceedings of the* 2003 *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, **1**, 16-23. https://doi.org/10.3115/1073445.1073448

[28] Zhao, S., Wang, H., Liu, T. and Li, S. (2008) Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. *Proceedings of ACL*-08: *HLT*, Columbus, OH, June 2008, 780-788.

[29] Quirk, C., Brockett, C. and Dolan, W. (2004) Monolingual Machine Translation for Paraphrase Generation. *Proceedings of the* 2004 *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, July 2014, 142-149.

[30] Zhao, S., Niu, C., Zhou, M., Liu, T. and Li, S. (2008) Combining Multiple Resources to Improve SMT-Based Paraphrasing Model. *Proceedings of ACL*-08: *HLT*, Columbus, OH, June 2008, 1021-1029.

[31] Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J. and Farri, O. (2016) Neural Paraphrase Generation with Stacked Residual LSTM Networks. *Proceedings of COLING* 2016, *the* 26*th International Conference on Computational Linguistics*:

Technical Papers, Osaka, Japan, 11-17 December 2016, 2923-2934.

[32] Li, Z., Jiang, X., Shang, L. and Li, H. (2017) Paraphrase Generation with Deep Reinforcement Learning. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 31 October-4 November 2018, 3865-3878. https://doi.org/10.18653/v1/D18-1421

[33] An, Z. and Liu, S. (2019) Towards Diverse Paraphrase Generation Using Multi-Class Wasserstein GAN. arXiv preprint arXiv:1909.13827.

[34] Gupta, A., Agarwal, A., Singh, P. and Rai, P. (2018) A Deep Generative Framework for Paraphrase Generation. 32*th AAAI Conference on Artificial Intelligence*, New Orleans, LA, 2-7 February 2018, 5149-5156.

[35] Galliers, J.R. and Jones, K.S. (1993) Evaluating Natural Language Processing Systems.

[36] Jones, K.S. and Galliers, J.R. (1996) Evaluating Natural Language Processing Systems : An Analysis and Review. *Lecture Notes in Computer Science*, **1083**, 336-338. https://doi.org/10.1007/BFb0027470

[37] Callison-Burch, C., Cohn, T. and Lapata, M. (2008) ParaMetric: An Automatic Evaluation Metric for Paraphrasing. *Proceedings of the* 22*nd International Conference on Computational Linguistics* (*Coling* 2008), Manchester, 18-22 August 2008, 97-104.

[38] Liu, C., Dahlmeier, D. and Ng, H.T. (2010) PEM: A Paraphrase Evaluation Metric Exploiting Parallel Texts. *Proceedings of the* 2010 *Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October 2010, 923-932.

[39] Chen, D.L. and Dolan, W.B. (2011) Collecting Highly Parallel Data for Paraphrase Evaluation. *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, **1**, 190-200.

[40] Nenkova, A., Passonneau, R. and McKeown, K. (2007) The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing*, **4**, 4.

[41] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, Philadelphia, PA, 7-12 July 2002, 311-318.

[42] Banerjee, S. and Lavie, A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, June 2005, 65-72.

[43] Mikolov, T., Yih, W. and Zweig, G. (2013) Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the* 2013 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Atlanta, GA, 9-14 June 2013, 746-751.

[44] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135-146. https://doi.org/10.1162/tacl_a_00051

[45] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. and Mikolov, T. (2018) Learning Word Vectors for 157 Languages. arXiv Preprint arXiv:1802.06893, 1-5.

[46] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. and Callison-Burch, C. (2015) PPDB 2.0: Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification. *Proceedings of the* 53*rd Annual Meet-*

ing of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, **2**, 425-430.
https://doi.org/10.3115/v1/P15-2070

[47] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. (2011) Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, **1**, 142-150.

[48] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv Preprint arXiv:1301.3781, 1-12.

[49] Bannard, C. and Callison-Burch, C. (2005) Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, 597-604.

[50] Ganitkevitch, J., Van Durme, B. and Callison-Burch, C. (2013) PPDB: The Paraphrase Database. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, June 2013, 758-764.