Evaluating Image-Inspired Poetry Generation

Chao-Chung Wu[1], Ruihua Song[⋆2][0000−0001−6036−9035], Tetsuya Sakai[3], Wen-Feng Cheng[1], Xing Xie[4], and Shou-De Lin[1]

[1] National Taiwan University, Taipei, Taiwan
r05922042@ntu.edu.tw,knowbee123@gmail.com,sdlin@csie.ntu.edu.tw
[2] Microsoft XiaIce, Beijing, China rsong@microsoft.com
[3] Waseda University, Tokyo, Japan tetsuyasakai@acm.org
[4] Microsoft Research Asia, Beijing, China xingx@microsoft.com

Abstract. Creative natural language generation, such as poetry generation, writing lyrics, and storytelling, is appealing but difficult to evaluate. We take the application of image-inspired poetry generation as a showcase and investigate two problems in evaluation: 1) how to evaluate the generated text when there are no ground truths, and 2) how to evaluate nondeterministic systems that output different texts given the same input image. Regarding the first problem, we first design a judgment tool to collect ratings of a few poems for comparison with the inspiring image shown to assessors. We then propose a novelty measurement that quantifies how different a generated text is compared to a known corpus. Regarding the second problem, we experiment with different strategies to approximate evaluating multiple trials of output poems. We also use a measure for quantifying the diversity of different texts generated in response to the same input image, and discuss their merits.

Keywords: Evaluation · poetry generation · natural language generation · AI-based creation · image.

1 Introduction

With the blossom of deep neural networks, some interesting studies on "creative artificial intelligence" (creative AI) have been reported, such as drawing a picture, composing a song, and generating a poem. Such tasks are attractive but also challenging. The biggest challenge posed by the research in creative AI is how to evaluate created content. Without a sound evaluation methodology, we cannot discuss scientific findings. While some initial studies on the evaluation of tasks related to creative AI have been reported (See Section 2), there remain many open problems, especially given the advent of neural models that can generate text.

In this paper, we take image-inspired poetry generation as a showcase to investigate some practical problems with evaluation. As Cheng *et al.* and Liu *et al.*[2, 17] described, image-inspired poetry generation is an application that takes a user's uploaded image as an input and generates a poem that is interesting to the user with the image content. In contrast to the well-known Image to Caption that requires a precise description of the image, an exemplary generated poem should have the following properties:

⋆ Ruihua is the corresponding author.

1. It is readable, i.e., each sentence is correct and sentences are logically coherent.
2. The content is related to the image. It is not necessarily relevant to all parts of the image, but relevant to some part(s).
3. It is novel. At least sentences are not in existing poems. It is more novel if fewer fragments are copied from elsewhere.

There are two major challenges in evaluating image-inspired poem generation. First, we need to evaluate the generated text even though there are no ground truths. As the goal of creative AI is to generate something novel, it may not be adequate for us to compare the generated text with a small set of ground truths or with texts from an existing corpus. Second, we evaluate nondeterministic systems, i.e., those that may output different texts given the same input image. As reported in Cheng *et al* [2], about 12 million poems have been generated from users as by August, 2018. In this kind of real application, different images may have the same set of tags. However, the users may find it boring if we always generate the same poem. While it is not difficult to devise nondeterministic neural generation models, e.g., Cheng *et al* [2] do not select the best candidate but one from n best results by taking a random factor into account in beam search, this poses a new challenge in evaluation.

As an initial investigation into the aforementioned challenges, we conduct experiments to evaluate image-to-poem methods based on neural models. First, we hire assessors to collect human labels for the generated poems, to use them as our gold standard. We find that the inter-assessor agreement doubles when an image is shown to the assessors as a context compared to when it is not. Second, we propose applying a simple novelty measure that quantifies how different a generated poem is from the training data as a complementary measure to ratings. Third, we address the problem of evaluating nondeterministic poetry generation systems by considering the diversity of the generated poems given the same input image. Our results indicate evaluating nondeterministic systems based on a single random trial may be a cost-effective evaluation method, i.e., assessing multiple times for each trial of nondeterministic system is exhausting, and the one-best evaluation of deterministic system also differs from the evaluation of a nondeterministic system. Fourth, we also propose a measure for quantifying the diversity of different texts generated in response to the same input image. Experiments indicate that diversity is complementary to novelty and human ratings, in particular for a large scale image-inspired poetry generation system.

2 Related Work

The growth of deep learning has generated great interest in natural language generation tasks, such as poetry generation and image to caption generation, but little work has been done on evaluation. Sparck Jones and Galliers [13] and Mellish and Dale [18] give overviews of existing evaluation methods, such as accuracy evaluation and fluency evaluation. They raise issues and problems, such as what should be measured and how to handle disagreement among human judges,

many of which have never been fully explored until now. For machine translation, Papineni *et al.* [20] propose an evaluation metric called Bilingual Evaluation Understudy (BLEU) that can automatically evaluate translation results with references based on the matching of n-grams. As it is efficient, inexpensive, and language independent, BLEU is widely adopted as a major measurement in machine translation. Some works like Stent *et al.* [24] make comparisons between several automatic evaluation metrics like BLEU score and F-measure, on different tasks, and point out some aspects which they omit, like the adequacy of the sentence. Galley *et al.* [7] propose ΔBLEU to allow a diverse range of possible output by introducing a weighted score for multi-reference BLEU. Hastie and Belz [10] focus on evaluating end-to-end NLG systems. However, most of these works focus on applying existing evaluation metrics to a more suitable task. With respect to AI based creation like storytelling, poetry generation and writing lyrics, the lack of ground-truth makes the BLEU score less suitable. In addition, there is an important feature that has been underlooked: creativity.

In terms of poetry writing, there are many generation tasks as mentioned in Colton *et al.* [4]; for either traditional or modern Chinese poetry, there are some works that propose poem generators (Hopkins and Kiela [12], Ghazvininejad *et al.* [8], He *et al.* [11], Zhang and Lapata [29], Yan [27], Wang *et al.* [26]), Cheng *et al.* [2] and Liu *et al.* [17]. For such tasks that require creativity, most of them use perplexity (PPL) for assessing training model capabilities and BLEU scores on testing as an automatic evaluation metric. However, PPL cannot guarantee good testing performance, and a lower PPL makes the model overfit to predict almost the same sentences given the same inputs, which is exemplary of a lack of creativity. Meanwhile, the BLEU score somehow cannot represent user favor as recent work by Devlin *et al.* [6] show that the BLEU score is not consistent with human ratings for image to caption generation. For evaluation not using BLEU, Ghazvininejad *et al.* [8] exploit human-machine collaboration and rating systems to improve and evaluate generated poetry. Hopkins and Kiela [12] propose intrinsic evaluations like examining rhythmic rules by phonetic error rate and extrinsic evaluations with indistinguishability studies between human and machine generated poetry. One of the image inspired poetry generation ,Liu *et al.* [17] also proposes to use visual-poetic embedding to calculate relevance score to consider coherence between image and poetry.

For creativity evaluation, Jordanous [14] conducts a survey on how creativity is evaluated and defined. She proposes the SPECS evaluation system including four key frameworks: person, product, process and environment are taken into consideration during evaluation. Zhu *et al.* [30] propose a set of quantified n-gram features combined with cognitive psychology features to represent the creativity of a single English sentence. Boden [1] makes the important distinction between H- (Historical) creativity (producing an idea/artifact that is wholly novel within the culture, not just new to its creator) and P- (Personal) creativity (producing an idea/artifact that is original as far as the creator is concerned, even though it might have been proposed or generated elsewhere and at an earlier time period). Ritchie [22, 23] defines two properties in assessing creativity: Novelty (to what

extent is the produced item dissimilar to existing examples of its genre?) and Quality (to what extent is the produced item a high quality example of its genre?). In our work, we take account of all the three aspects. We propose using human ratings to measure quality, novelty to measure H-creativity, and diversity to measure P-creativity.

Studies most relevant to ours are those on evaluating poetry generation. Lamb *et al.* [15] propose evaluating a template-based poetry generator, PoeTryMe (Oliveira [19]), and evaluate generated poetry with intra-class judges correlation, significant testing between judges, and analysis on factors of quality. Under the same generator framework, Oliveira [9] proposes a multilingual extension and the evaluation of the generator, which evaluates the poetic, structure, and topicality features of multilingual generated poems with ROUGE (Lin [16]), Pointwise Mutual Information (PMI) (Church and Hanks [3]), and other such methods. Velde *et al.* [25] propose a semantic association for evaluating creativity, which extracts creative words provided by human judges and analyzes the creative level and aspects of the words. For evaluation on an RNN based generator, besides the BLEU, Potash *et al.* [21] propose an LSTM rap lyrics generator and evaluates artistic style by similarity of lyric style and rhyme density. Although many studies on evaluation have been reported, most of them evaluate template/corpus based generators. As we are evaluating an RNN based generator, some traits of information in the generation of a neural network can be evaluated by controlling inputs. We are able to measure how diverse a generator can be when given the same input, which is rarely discussed. In this paper, we are evaluating RNN based generators such as what Cheng *et al.* and Liu *et al.* [2, 17] proposed.

3 Evaluation without Ground Truths

3.1 Collecting Human Ratings

Although it is costly, the best way to evaluate creative AI is leveraging human beings. Still we need to carefully design an annotation tool with guidelines and manage the process for collecting reliable ratings that are consistent with user satisfaction.

Annotation Tool Design Collecting reliable human assessments is an important step. We do not choose a design that shows an image and a poem each time and asks for a rating from assessors because such ratings are not stable for comparing poem quality, as assessors may change their standards unconsciously. A-B testing in search evaluation is better for comparing two methods, in particular when user satisfaction involves many factors that cannot be explicitly described or weighted. The disadvantage of A-B testing is two-fold: 1) the workload and cost dramatically increase when we would like to compare more than two methods because we may have to evaluate each pair of methods. 2) it is not adequate for us to learn the absolute level of user satisfaction that is helpful

to track changes among a series of approaches, as we only know the preference between two methods.

Through some trials, we design the interface of an annotation tool that takes into account of both absolute judgment and relative judgments. As shown in Figure. 1, we present an image at the top and the poems generated by different methods for comparison side by side below the image. We randomize the order of methods for each image and mask the methods from the assessors, thus removing biases. For each poem, we ask assessors to give a rating from one to five after comparing the poems. An assessor can easily read and compare all poems before rating, and thus his/her scores can provide meaningful information on the relative ordering of poems. At the same time, we give detailed guidelines on the five levels of ratings and thus we can collect the ratings that are comparable between images and methods.

Annotation Guidelines Specifically, we ask assessors to consider the following factors when they judge a poem:

1. Whether each sentence uses correct diction and syntax;
2. Whether a poem is related to the image;
3. Whether sentences of a poem are logically coherent;
4. Whether some part of the poem is imaginative and/or moving.

When all sentences are understandable, i.e., conditions (1) and (2) are satisfied, we recommend that assessors give a rating of 3. Above that, assessors can give a rating of 4 if the poem is logically coherent, i.e., condition (3) is also satisfied. A rating of 5 corresponds to cases where the poem has some highlights, i.e., condition (4) is satisfied further. On the other hand, if a poem is not related to any part of the image or some sentences have incorrect words, collocation or grammar, assessors can subtract one or two points from the 3 rating. Usually, if only one sentence is not understandable, we suggest they give a rating of 2; if more than one are not understandable or worse, they can give a rating of 1.

3.2 Novelty

As our task is a kind of creative language generation, the poem should not be composed entirely of copies from different parts of existing poems. For example, the repeat fragment "city is too ashamed to face the countryside" comes from the poem shown in Figure. **??**. Thus, generated sentences like "This *every* city is too ashamed to face the countryside" is not considered very novel. It would be more novel if fewer fragments overlapped. We propose using Novelty to measure how culturally novel a created sentence/poem is to existing poems, denoted here by a training corpus.

First we calculate $V_{k,i}$ as the ratio of k-grams that are novel to the training data in sentence i:

$$V_{k,i} = \frac{\#(novel\ k\text{-}grams\ in\ i)}{\#(k\text{-}grams\ in\ i)}. \quad (1)$$

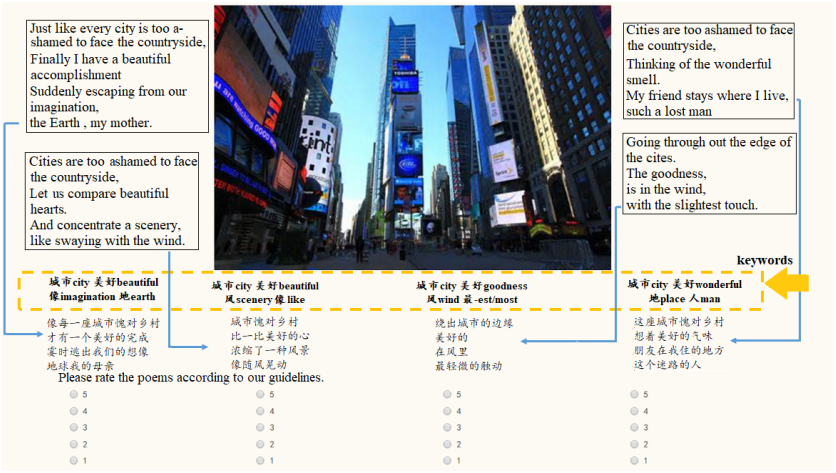

Fig. 1. The human assessment tool is designed to capture both the relative judgments among methods and absolute ratings. Since there is a dividing-sentence trait for Chinese poetry, in our English translation, every comma or period indicates the end of one single Chinese poetry line.

We then calculate the novelty of a sentence i as follows:

$$novelty_i = \frac{\sum_{k=3}^{n} V_{k,i}}{n-2}, \text{for } n = \min(8, L_i). \tag{2}$$

where, L_i is the length of the sentence. $n = min(8, L_i)$ can guarantee the denominator of $V_{k,i}$ is larger than zero.

For a poem that is composed of N sentences, the novelty of the poem is calculated as follows:

$$novelty = \frac{\sum_{i=1}^{N} novelty_i}{N}, \text{for } N = 4. \tag{3}$$

When a whole poem comes from a training corpus, it is not novel at all. The corresponding novelty score is 0 because no k-grams are new. On the other hand , when a poem is entirely new in terms of all tri-grams, the novelty score is 1, meaning it is extremely novel.

Finally, we use the mean novelty for a set of poems generated for our test image set.

3.3 Experiments

Does an image matter? In such a creative generation task, human evaluation is subjective. We do not think it is reasonable to require as high a level of agreement between assessors in poetry generation as that in information retrieval. However, we are curious whether the disagreement between assessors in

Table 1. Pearson correlations between human ratings, novelty, and diversity

Person Correlation	Rating	Novelty	Diversity
Rating	1	-0.59	-0.65
Novelty	-	1	0.19
Diversity	-	-	1

our image-inspired poetry generation application is similar to that in poetry generation without an image. Thus, we design a user study to investigate the question. We first invite three human assessors to rate generated poems via the tool in Figure. 1 but without showing the image. There are fifty pages corresponding to the fifty images that inspire the generation. On each page, we show four poems that are generated by four different methods. They are anonymised and their order is randomized. The guidelines are those described in Section 3.1 except for the second criteria on relatedness between poems and the image. After the first round, we ask assessors to take a rest for thirty minutes to reduce the assessors' impression on the previous poems. Next, we ask the assessors to rate the fifty pages of poems again but with images shown as in Figure. 1, and according to the full guidelines.

Once we collect the rating, the agreement between assessors' rating with or without images is calculated by Kendall tau-b correlation coefficient between two assessors. Then we further average Kendall tau scores over the fifty images and three pairs of assessors. Then we can observe whether images provide more consistency for users ratings. Our results show that with images the coefficient is as high as 0.27; while without image, the coefficient drops to 0.11. Such results indicate that although the ratings on poetry generation are still subjective, with images provided in the evaluation, assessors can more easily get agreement on the rating order of poems. The reasons may be that assessors do not simply rates poems based on the word content but considering the context of image. The image context can make assessors better understand poems' meanings and rate them.

Human Ratings vs Novelty We collect the human ratings and calculate novelty for the eight methods for comparison. We invite 28 subjects to participate in our manual evaluation. Our subjects include 16 males and 12 females. Their average age is 23 with a range from 18 to 30. To reduce the bias of users and order, we apply the Latin Square methodology to arrange the labeling task to 28 subjects. The results are shown in Figure. 2. We also calculate the Pearson correlation for each pair as shown in Table 1. The correlation between Rating and Novelty is -0.59. Over methods m_1, m_2, and m_3, we observe that the higher the human rating is, the lower the novelty is. Methods m_5 and m_6 are also following this trend. This can be explained in the way that m_1, m_2, and m_6 generate too many new words or new combination with sacrifice of correctness. compared to m_4, m_3 and m_5 can improve both Rating and Novelty. Thus, the two measurements together can help us find the truly better methods.

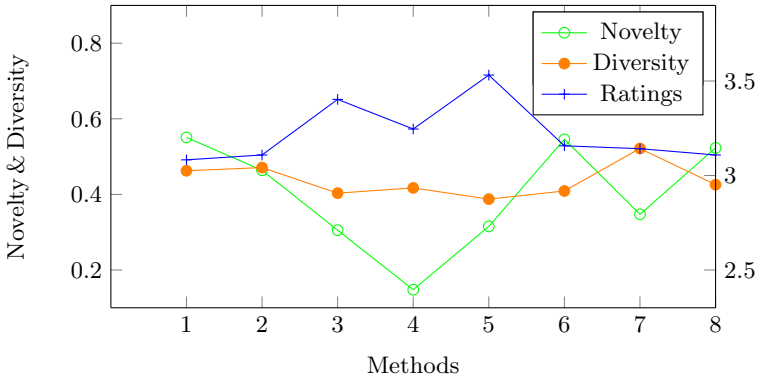

Fig. 2. Evaluation results of eight methods in terms of human rating, novelty and diversity.

4 Evaluation of Nondeterministic Systems

4.1 Diversity Measure

Similar to diversity defined in Deng *et al.*[5] and Zhang and Hurley [28], we leverage the Jaccard Distance of sets to calculate diversity between a set of the i-th sentences $s_1, s_2, ..., s_M$ of M poems:

$$diversity_i = \frac{\sum_{k=1}^{n} D_{k,i}}{n-2}, for\ n = 8. \tag{4}$$

where $D_{k,i}$ is defined as follows:

$$D_{k,i} = \frac{|s_1^k \oplus s_2^k \ldots \oplus s_M^k|}{|s_1^k \bigcup s_2^k \ldots \bigcup s_M^k|}. \tag{5}$$

where, $\oplus$ is defined as the XOR set operator. s_j^k is the set of k-grams in sentence s_j. If k is larger than the length of s_j^k, the set becomes null.

The diversity of poems is the average of all K sentences in a poem:

$$diversity = \frac{\sum_{i=1}^{K} diversity_i}{K}. \tag{6}$$

4.2 Experiments

Deterministic vs. Nondeterministic Like the applications of image to caption or machine translation, we can return the best result in beam search, which is a deterministic one-best result. How different is the one-best result from the average human ratings for three trials of a nondeterministic system?

In our user study as described in Section 3.3, for an image, each method generates four poems, in which three are generated with a random among n

Table 2. Correlations between the ratings of one-best, one trial, and three trials

Pearson Correlation	One-Best	Average-Random	One-Random
One-Best	1	0.473	0.387
Average-Random	-	1	0.925
One-Random	-	-	1

best approach in beam search (as our system is nondeterministic, it is better to evaluate a method over several trials) and one is the one-best result in beam search (this is designed for an experiment in Section 4.2). To compare two different poem generation methods, we present all eight poems generated by the two methods for an image. The interface as shown in Fig. 1 will have a horizontal scroll-bar when the number of poems is larger than four. As a result, we collect human ratings for one result generated by a one-best strategy and three results by random sampling strategy.

We can regard the average ratings over the three results by random sampling, a.k.a., Average-Random, as the ground truth, since what the users actually experience is a nondeterministic system. Then we calculate the Pearson correlations between human labels of the one-best result (a.k.a., One-Best), human labels of one trial of random (a.k.a., One-Random), and the average human labels of three trials of random. As we have three random results, we calculate the Pearson correlation between each of them and the One-Best and then average the three correlations to get the correlation of One-Random and One-Best. In the same way, we calculate the correlation of One-Random and Average-Random. Results are shown in Table 2.

We have a few interesting findings from Table 2. First, it can be observed that the correlations between the ratings for One-Best and those for the two sets of Random results are not high (0.387-0.473). This means that our nondeterministic system behaves differently from the traditional approach that relies on the one-best result from beam search. Second, and more importantly, it can be observed that the correlation between Average-Random and One-Random is as high as 0.925. This suggests that, given a limited budget, observing just one random result per input image may suffice to evaluate the entire nondeterministic system.

Diversity of Methods In this experiment, we use the models from Cheng *et al.* [2] as a poetry generator. We use the generated results to calculate diversity, where the number of poems is $M = 3$. The mean diversities of the methods are also shown in Fig. 2. We calculate the Pearson correlation between Diversity and the other three measurements respectively as shown in Table 1.

The correlation between Diversity and Rating is -0.65. This indicates that higher ratings may be achieved by sacrificing diversity to some extent. For example, the method m7 is worse than m5 in terms of ratings, but it achieves much better diversity. Novelty and Diversity yield positive correlations as low as 0.19. This suggests that Diversity and Novelty are different. For example, m_7 is better than m_1 in terms of both Ratings and Diversity, but it is worse than

m_1 in terms of Novelty. Such a phenomenon is possible when m_1 generates more different sentences, which may be less readable but new to the training corpus.

5 Conclusion

In this paper, we investigate the fundamental problems with evaluation of deep neural network based methods for image-inspired poetry generation. We design an annotation tool to collect human ratings while keeping relative orders between methods. Our user study results indicate that showing an image can double the Kendall tau of a poem ranking between different assessors from 0.11 to 0.27. Moreover, we find that human ratings cannot measure the novelty of created poems to existing poems for training. Hence, we use novelty as a complementary measure to human ratings. In a real application with large-scale user requests, our system is designed to be nondeterministic so that diverse poems can be generated at different times in response to the same image. Our experiments show that the human ratings of one-best deterministic results have correlation as low as 0.473 with human ratings over three trials of our nondeterministic system; whereas, the correlation between the human ratings for one trail of a nondeterministic system and three trials is as high as 0.925. This suggests that evaluating nondeterministic systems based on a single random trial may be a cost-effective evaluation method. Finally, we find that diversity is also necessary to measure non-deterministic systems in additional to Rating and Novelty.

As for limitations of our work, Novelty is really to do with semantics, but we only look at overlaps of surface strings to evaluate novelty. We would like to conduct more research on this topic. In addition, we plan to extend our evaluation methodology to other creative AI tasks, such as writing lyrics or a song.

References

1. Boden, M.A.: The Creative Mind: Myths and Mechanisms. Basic Books, Inc., New York, NY, USA (1991)
2. Cheng, W.F., Wu, C.C., Song, R., Fu, J., Xie, X., Nie, J.Y.: Image inspired poetry generation in xiaoice. CoRR **abs/1808.03090** (2018)
3. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (Mar 1990), http://dl.acm.org/citation.cfm?id=89086.89095
4. Colton, S., Goodwin, J., Veale, T.: Full-face poetry generation. In: Proceedings of the Third International Conference on Computational Creativity (ICCC'12) (2012)
5. Deng, F., Siersdorfer, S., Zerr, S.: Efficient jaccard-based diversity analysis of large document collections. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 1402–1411. CIKM '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2396761.2398445, http://doi.acm.org/10.1145/2396761.2398445

6. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)(ACL'17) (2015)
7. Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., Dolan, B.: deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 445–450. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-2073
8. Ghazvininejad, M., Shi, X., Priyadarshi, J., Knight, K.: Hafez: an interactive poetry generation system. In: Proceedings of ACL 2017, System Demonstrations. pp. 43–48. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-4008, http://aclanthology.coli.uni-saarland.de/pdf/P/P17/P17-4008.pdf
9. Goncalo Oliveira, H., Hervas, R., Diaz, A., Gervas, P.: Multilingual extension and evaluation of a poetry generator. Natural Language Engineering **23**(6), 929–967 (2017). https://doi.org/10.1017/S1351324917000171
10. Hastie, H., Belz, A.: A comparative evaluation methodology for nlg in interactive systems. In: Calzolari, N. (ed.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (2014)
11. He, J., Jiang, L., Ming, Z.: Generating chinese couplets using a statistical mt approach. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. pp. 377–384. COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), http://dl.acm.org/citation.cfm?id=1599081.1599129
12. Hopkins, J., Kiela, D.: Automatically generating rhythmic verse with neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 168–178. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1016, http://aclanthology.coli.uni-saarland.de/pdf/P/P17/P17-1016.pdf
13. Jones, K.S., Galliers, J.R.: Evaluating Natural Language Processing Systems: An Analysis and Review. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1996)
14. Jordanous, A.: A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation **4**(3), 246–279 (Sep 2012). https://doi.org/10.1007/s12559-012-9156-1, https://doi.org/10.1007/s12559-012-9156-1
15. Lamb, C., Brown, D., Clarke, C.: Evaluating digital poetry: Insights from the CAT. In: Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Sony CSL, Sony CSL, Paris, France (2016), http://www.computationalcreativity.net/iccc2016/wp-content/uploads/2016/01/Evaluating-digital-poetry.pdf
16. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (July 2004), https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/
17. Liu, B., Fu, J., Kato, M.P., Yoshikawa, M.: Beyond narrative description: Generating poetry from images by multi-adversarial training. In: Proceedings of

the 26th ACM International Conference on Multimedia. pp. 783–791. MM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3240508.3240587, http://doi.acm.org/10.1145/3240508.3240587
18. Mellish, C., Dale, R.: Evaluation in the context of natural language generation. Computer Speech & Language **12**(4), 349 – 373 (1998). https://doi.org/http://dx.doi.org/10.1006/csla.1998.0106, http://www.sciencedirect.com/science/article/pii/S0885230898901061
19. Oliveira, H.G.: Poetryme: a versatile platform for poetry generation. Computational Creativity, Concept Invention, and General Intelligence **1**, 21 (2012)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02). pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1073083.1073135, https://doi.org/10.3115/1073083.1073135
21. Potash, P., Romanov, A., Rumshisky, A.: Evaluating Creative Language Generation: The Case of Rap Lyric Ghostwriting. ArXiv e-prints (Dec 2016)
22. Ritchie, G.: Assessing creativity. In: Proceedings of the AISB01 Symposium on Artificial Intelligence and Creativity in Arts and Science. pp. 3–11 (2001)
23. Ritchie, G.: Some empirical criteria for attributing creativity to a computer program. Minds and Machines **17**(1), 67–99 (Mar 2007). https://doi.org/10.1007/s11023-007-9066-2, http://dx.doi.org/10.1007/s11023-007-9066-2
24. Stent, A., Marge, M., Singhai, M.: Evaluating Evaluation Methods for Generation in the Presence of Variation, pp. 341–351. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
25. van der Velde, F., van der Velde, F., Wolf, R., Schmettow, M., Nazareth, D.: A Semantic Map for Evaluating Creativity, pp. 94–101. WordPress (6 2015), open access
26. Wang, Q., Luo, T., Wang, D.: Can machine generate traditional chinese poetry? a feigenbaum test. In: BICS (2016)
27. Yan, R.: i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In: Kambhampati, S. (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 2238–2244. IJCAI/AAAI Press (2016), http://www.ijcai.org/Abstract/16/319
28. Zhang, M., Hurley, N.: Avoiding monotony: Improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM Conference on Recommender Systems. pp. 123–130. RecSys '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1454008.1454030, http://doi.acm.org/10.1145/1454008.1454030
29. Zhang, X., Lapata, M.: Chinese Poetry Generation with Recurrent Neural Networks, pp. 670–680. Association for Computational Linguistics (10 2014)
30. Zhu, X., Xu, Z., Khot, T.: How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives. In: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. pp. 87–93. CALC '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1642011.1642023