Automatic Arabic Poem Generation with GPT-2

Mohamed El Ghaly Beheitt[1][a] and Moez Ben Haj Hmida[2][b]
[1] *LIPAH-FST Laboratory , Faculty of Sciences of Tunis, University of Tunis El Manar , Tunis, Tunisia*
[2] *National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia*

Keywords: Deep Learning, Transformer, Natural Language Processing, GPT-2, Arabic Poem.

Abstract: Automatically generating poetry by computers is a challenging topic that requires the use of advanced deep learning techniques. While much attention has been given to English and Chinese poem generation, there are few significant efforts considering other languages. Generating poems in Arabic is a difficult task due to the complexity of the Arabic language grammatical structure. In this paper, we investigate the feasibility of training generative pre-trained language model GPT-2 to generate Arabic poems. The results of the experiments, which included the BLEU score as well as human assessments, confirmed the effectiveness of our proposed model. Both automatic and human evaluations show that our proposed model outperforms existing models in generating Arabic poetry.

1 INTRODUCTION

Poetry is a unique and essential cultural treasure that dates back thousands of years in human history. Poetry popularity may be seen in many facets of daily life, such as expressing personal emotion, political opinions, or delivering messages at celebratory events. Writing poetry, as a fascinating art form, is an appealing problem that Artificial Intelligence (AI) researchers are interested in (Yan et al., 2013; Das and Gambäck, 2014; Oliveira and Cardoso, 2015; Ghazvininejad et al., 2016; Ghazvininejad et al., 2017; Singh et al., 2017; Xu et al., 2018; Zugarini et al., 2019; Van de Cruys, 2020). This interest is motivated by the fact that poetry generation is an application of Natural Language Generation (NLG).

NLG is a challenging topic that has attracted the interest of the Natural Language Processing (NLP) community (Subramanian et al., 2017). Traditionally, rule-based methods (Zhou et al., 2010) and statistical machine translation models (He et al., 2012) are recommended for this task. Deep neural networks have recently been used to create fluent and natural poetry (Wang et al., 2016a; Zhang et al., 2017). Although these models appear promising, they are constrained in many ways. For example, past research often fails to maintain theme coherence (Wang et al., 2016c; Yang et al., 2017) and increase term variety (Zhang et al., 2017), both of which are essential properties of poems. Compared to efforts performed in English and Chinese for NLG, Arabic applications of deep learning for NLG, particularly Arabic poem generation, are still limited.

Arabic poetry is the earliest form of Arabic literature. In the Arabic literary tradition, poetry has reflected the deepest sense of Arab self-identity, of communal history, and of aspirations for the future. Arabic poetry is often divided into two categories: classical poetry and modern poetry (also named free poetry). Consequently, any poetry composed in the classical form is referred to as ”traditional poetry” since it adheres to the conventional form and structure. It is sometimes referred to as ”vertical poetry” because of the vertical parallel construction of its two parts known as hemistichs. Modern poetry, on the other hand, varied from traditional poetry in terms of style, structure, rhyme, and subjects. Arabic poetry is defined as rhymed, metered speech. According to some definitions, Arabic poetry is an eloquent, rhyming speech, at the end of which there is a rhyme and a musical rhythm; it is often expressed in rhetorical images, using imagination that aims to evoke conscience and feeling. Therefore, the Arabs are instinctively led to love poetry and formulate it.

The Arabs did not know the meters (rhythmic structure of a verse) of poetry by learning specific laws and systems from the beginning. Rather, they organized the poems by their nature according to

[a] https://orcid.org/0000-0001-8414-7293
[b] https://orcid.org/0000-0001-7682-4549

Beheitt, M. and Ben Haj Hmida, M.
Automatic Arabic Poem Generation with GPT-2.
DOI: 10.5220/0010847100003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 2*, pages 366-374
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Table 1: Example of verses from Arabic poems.

(Second hemistich) الشطر الثاني	**(First hemistich)** الشطر الأول
الجود يفقر و الإقدام قتال The generous are poor, and courage kills its own	لولا المشقة ساد الناس كلهم Without hardship everyone would prevail
و السيف و الرمح و القرطاس و القلم As do the sword, the spear, the scripture and the pen	الخيل و اللّيل و البيداء تعرفني The steed, the night and the desert all know me

what was dictated to them by the chanting format. Table 1 shows examples of verses from Arabic poems by the famous Arab poet Abu al-Tayyib Ahmad ibn Al-Husayn Al-Mutanabbi.

In this paper, we describe the process of pre-training a customized Open-AI Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019) for the Arabic language. We trained the model on over 1 Million Arabic news. Then, we fine-tuned our pre-trained model Arabic poem generation and compared the model against state of the art models.

The remainder of this paper is organized as follows. Section 2 reviews some recent related works. Section 3 goes through our proposed approach for generating Arabic poems. Section 4 presents experiments, evaluation, and obtained results, followed by conclusions and future work in Section 5.

2 RELATED WORK

Poetry generation is arguably the toughest of the text generation subtasks. Since the poem must be generated in an elegant manner and ideally following a particular structure.

Automated poetry generation has been a common research topic over the last few decades. (Wang et al., 2016b), used an attention-based LSTM (Long-Short Term Memory) model for the iambics generation of Chinese songs. The model asks users to enter a first line for the poem, and then the model generates the other lines. To learn the vector representations of the words, authors used Word2Vec (Guo et al., 2018). The results evaluated on the basis of subjective and automatic performance measures show a better quality of the proposed model over statistical machine translation (SMT)(He et al., 2012) and the RNN linguistic model (RNNLM)(Mikolov et al., 2010).

In (Yan, 2016) Yan proposed a polishing framework based on recurrent neural network (RNN) to generated Chinese poems. This framework encodes user intents and generates poems via a sequential generation. This work uses a polishing schema to refine poem composition until a well-formed one is generated.

Similarly, (Yi et al., 2018) presented a salient-clue mechanism. Their model automatically selects the most salient characters from the last generated lines and uses the selected characters as a theme clue for generating the next lines of the poem. This mechanism enhances the meaning and coherence of generated poems in Chinese language.

Yi et al. (Yi et al., 2017) based their work on a sequence-to-sequence model (Cho et al., 2014) to generate Chinese poems. They built an encoder-decoder framework based on a bi-directional recurrent neural network (Bi-RNN) with an attention mechanism.

In (Wei et al., 2018) authors attempted to solve the style problem in Chinese poetry by proposing Poet-based approach. The proposed method is divided into two stages. First, they capture poetic style embedding by modeling poems and high-level abstractions of poetic style in a Poetic Style Model. Second, they sequentially generate each line with a modified RNN encoder-decoder. Authors discovered that satisfactory results could be obtained with enough training data.

Lau et al. (Lau et al., 2018) developed a model for generating English quatrains (Shakespeare-like sonnets). To generate quatrains, authors used a joint architecture of three neural networks that capture the language, rhyme, and meter. They also used crowdsourcing and experts judgments to assess the quality of generated quatrains. Their crowdsourcing and expert evaluations indicated that the produced poems followed the sonnet structure but lacked readability and coherence.

Talafha and Rekabdar (Talafha and Rekabdar, 2019a; Talafha and Rekabdar, 2019b), are among the first researchers to use deep learning to generate Arabic poems. In (Talafha and Rekabdar, 2019b) they proposed to generate Arabic poetry using two models, a Bi-GRU (Bi-directional Gated Recurrent Unit) model for composing the first line of the poem and a modified Bi-GRU encoder-decoder model with hierarchical neural attention for producing other lines of the poem. They also proposed in (Talafha and Rekabdar, 2019a) a poetry generation model (Phonetic CNN subword embeddings) with expanded phonetic and semantic embeddings, which are concatenated embeddings that provide information on the phonetics of each word as well as its vectorized word representation. In both works, authors used BLEU scores and human evaluation to evaluate the generated poems. According to human evaluation, their models generated high-quality poems in terms of coherence, fluency, meaning, and poeticness.

Bena and Kalita (Bena and Kalita, 2020) proposed a new method to generate poems in English. They fine-tuned a pre-trained language model GPT-2 (Radford et al., 2019) to generate poems that express and elicit emotion in readers, as well as poems that use dream language, which is known as dream poetry. They classified emotion poems and dream text to influence automatic natural language generation to create poetry. To accomplish this task, they used a word-level emotion lexicon to create a meaning for emotion-eliciting text, which was then used to train separate GPT-2 models. To teach the language of poems to the network, authors pre-trained the OpenAI-released GPT-2 model on a corpus of first-person dream descriptions. Then, they fine-tuned the obtained model on a dataset of 20,000 dreams.

GPT-2 demonstrated high performance on NLP tasks (Radford et al., 2019) and on the generation of English poems (Bena and Kalita, 2020).

Recently, (Hakami et al., 2021) investigated the GPT-2 model in generating Arabic poems. Authors fine-tuned GPT-2 model pre-trained on English corpora. The fine-tuning dataset consisted of 34,466 Arabic verses manually collected from the aldiwan website[1]. The resulting GPT-2 model performed poorly in terms of BLEU-1 score (0.56) and human evaluation (0.5 in meaning and coherence).

The success of GPT-2-based English poem generation represents the main motivation of our work. In the following we detail how we built a GPT-2-based model for automatic Arabic poem generation.

3 PROPOSED APPROACH

3.1 Model Architecture

OpenAI developed an unsupervised transformer-based generative language model called GPT-2 (Generative Pre-trained Transformer 2) (Radford et al., 2019). The language model is a machine learning model that uses probability distributions to predict the next word of a given sentence. Language models use unsupervised methods to develop a lot of features that represent rules of spelling and grammar. Unsupervised learning methods consider the patterns in a set of data rather than trying to identify a relation between data. To predict the next word, GPT-2 was trained on a large corpus (WebText) containing 40 GB of text (Radford et al., 2019). To perform encoding, GPT-2 makes use of BPE (byte-pair encoding) (Sennrich et al., 2015). BPE encoding is a sub-word encoding, that is between character level and word level.

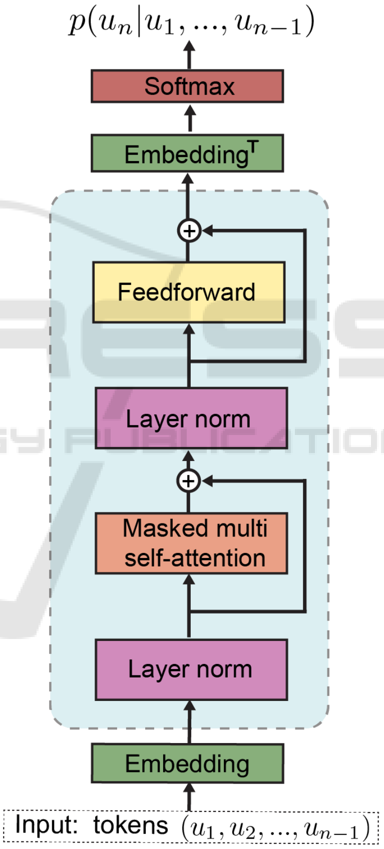
$p(u_n|u_1, ..., u_{n-1})$
Softmax
Embedding^T
Feedforward
Layer norm
Masked multi self-attention
Layer norm
Embedding
Input: tokens $(u_1, u_2, ..., u_{n-1})$

Figure 1: GPT-2 architecture,(Heilbron et al., 2019).

GPT-2 architecture has proven to model the English language and has obtained state-of-art tasks such as summarization, machine translation, and question

[1] https://www.aldiwan.net

answering. This model has four versions: extra-large version (1.5 billion parameters), a large version (762 million parameters), a medium version (345 million parameters), and a small version (117 million parameters). In this paper, we use the small version of GPT-2 model for the task of Arabic poetry generation, due to our limited compute capacity. As shown in Figure 1, the GPT-2 architecture is quite similar to the architecture of the decoder-only transformer (Vaswani et al., 2017). Input tokens are transferred through a token embedding matrix in the network. The activities are then routed through a stack of Decoder blocks, which includes a multi-headed self-attention layer, a position-wise feedforward layer, and a normalization layer.

3.2 Data Processing

The first step consists of collecting data to train our GPT-2 model. Pre-training and fine-tuning are the two stages of the training process. We used the two publicly available corpora Khaleej-2004 (Abbas and Smaili, 2005) and Watan-2004 (Abbas et al., 2011) to pre-train our model. Khaleej-2004 is an MSA (Modern Standard Arabic) corpus that was collected from thousands of articles downloaded from Akhbar Al Khaleej, an online newspaper. The corpus contains 5,690 documents, totaling more than 2 million words. The Watan-2004 is an also MSA corpus that composed of nearly 20,000 documents that correspond to more than 9 million words. To fine-tune our model, we used Arabic poetry dataset [2] that was scrapped entirely from aldiwan website[3]. There are 55K poems for over 540 poets from 9 different eras in the Arabic poetry dataset. Table 2 summarizes the datasets we use for training regarding the number of words and unique words for each dataset.

Table 2: Statistics of the used datasets.

Dataset	#Words	#Unique Words
Khaleej-2004	2.482K	122K
Watan-2004	9.813K	291K
Total pre-training	12.229K	413K
Arabic poetry	6.933K	2.060K

3.3 Training

3.3.1 Pre-training

To generate Arabic poems, we have to pre-train our model on the Arabic language. We got the pre-trained GPT-2 Tokenizer and Model from the Transformers Library (Hugging Face[4]). This Library provided us with the tokenizer structure we need as well as pre-trained model weights, rather than starting with random values, we started training our GPT-2 model in Arabic with weights that have already been trained in the English language. We trained a BPE tokenizer on the Arabic corpus using the Tokenizers Library (Hugging Face), which gave us the vocabulary files (vocabulary size 50K tokens) in Arabic of our GPT-2 tokenizer. We pre-trained our GPT-2 model on Google Colab with the Khaleej-2004 and Watan-2004 corpora. We trained it in NVIDIA Tesla T4 (16 GB) GPU for 25 epochs. The pre-training took about 32 hours. The resulting model in this pre-training stage can be fine-tuned to perform NLP Arabic tasks.

3.3.2 Fine-tuning

For our downstream task of Arabic poetry generation, we used the GPT-2 model pre-trained in the Arabic language in the fine-tuning process. By training this model on poem text from the Arabic poetry dataset, we are able to generate Arabic poems. We used the same GPU using in the pre-training stage for 6 epochs to fine-tune our model. The fine-tuning took 12 hours.

Table 3: Hyperparameters used for the poem generation.

Hyperparameter	Pre-training	Fine-tuning
Max sequence length	1024	1024
Batch size	24	24
Learning Rate	3e-05	3e-05
# Epochs	25	6

3.4 Poem Generation

Since the natural language generation is based on the notion that the probability distribution of a word sequence can be modeled in terms of the conditional probability of next word distributions (Bengio et al., 2003):

$$P(w_1, ..., w_T) = \prod_{t=1}^{T} P(w_t | w_1, ..., w_{t-1}) \qquad (1)$$

[2] https://www.kaggle.com/ahmedabelal/arabic-poetry

[3] https://www.aldiwan.net

[4] https://huggingface.co/

Table 4: BLEU Comparison.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Vanilla	0.0211	0.0199	0	0
LSTM	0.1522	0.1124	0.0081	0.0013
GRU	0.1512	0.1139	0.0084	0.0021
RNN EncoderDecoder (without attention)	0.2513	0.1539	0.0740	0.0510
RNN EncoderDecoder (with attention)	0.3010	0.2110	0.0911	0.0801
(Talafha and Rekabdar, 2019b) Model	0.4122	0.3144	0.204	0.1092
(Talafha and Rekabdar, 2019a) Model	0.5301	0.4010	0.3001	0.1500
GPT-2	0.8739	0.5369	0.3230	0.1871

Table 5: Human evaluation.

Models	Fluency	Coherence	Meaning	Poeticness
Vanilla	0.1	0.8	0.7	0
LSTM	0.3	0.9	0.8	0.1
GRU	0.3	1	1	0.2
RNN Encoder- Decoder (without attention)	2	1.5	2.4	0.3
RNN Encoder- Decoder (attention)	2.3	2.5	2.7	0.4
(Talafha and Rekabdar, 2019b) Model	2.1	3.2	3.5	0.9
(Talafha and Rekabdar, 2019a) Model	2.7	3.3	3.6	2.5
GPT-2 Model	2.8	2.6	2.6	3.4

Table 6: Example verses from poems generated by our model and poems generated by (Talafha and Rekabdar, 2019b) Model.

Model	The generated verses in Arabic	The generated verses in English
GPT-2 Model	وأبكي دموعا قد أضرت مدامعي وأهوى إليها أن تكون المداويا	And I shed tears that have damaged my eyes, and I want her to be the healer.
	وكانت على وجهي الربوع إلى منى تذكرني منها الدموع المواليا	Before my eyes, all the places are spread until Mina, and my tears point out the dearest amongst them .
	كان في الترحال عقد صبابة لا تستكن صبابة لعناني	He was untamed as love itself, for love can never be subjugted.
	على أحمد المختار خير الورى له مولاة أهل الغيب في ما يسره	To Ahmed, the chosen one, best of all, the unseen pledge allegiance and do he pleases.
	إن لم يكن لك شيء غير منتظر فما لها غير آمال بمفتون	If all you possess is certainly attainable, all she grasps are promises of an admirer.
(Talafha and Rekabdar, 2019b) Model	الى عرفات ارض رسول للّه محبه سلام اللّه على حجاج بيت للّه	To my beloved Arafat Allahs messengers land, peace be upon the pilgrims of Allahs house.
	ارض الحجاز يشدني الحنين لساكنها هنا النور قد أتينا يا رسول للّه	AL-Hijaz land, missing you is what attracts me to your people, Oh Allah's messenger, here is the light that brought us.
	جبريل الآمين عليك سلام اللّه في عرفات في بلاد طهرها للّه	Gabriel, peace upon you, in Arafat the purest land of Allah.
	يا كعبة الرحمن في مكة اهواها شوقا الى البيت الحرام اذوب	Oh, Al-Rahman's house in Mecca, how much I love you what made us cone is missing you.

Our approach generates the two verses of a poem. We used two sampling methods Top-K (Fan et al., 2018) and Top-p (nucleus) (Holtzman et al., 2019) to sample from the distribution in this probabilistic form of language modeling. Top-K sampling filters the K most likely next words and redistributes the probability mass of just those K next words. Top-p sampling selects the smallest possible set of words whose total likelihood exceeds the probability p, this set of words is then redistributed with the probability mass. We used the combination of Top-K and Top-p sampling strategies with K=40 and p=0.92 to generate diverse poems. All experiments are done with the values of hyper-parameters presented in Table 3.

4 EXPERIMENTS

4.1 Automatic Evaluation

We adopt BLEU (Bilingual Evaluation Understudy) scores (Papineni et al., 2002) to automatically evaluate the poems generated by the GPT-2 model. The BLEU is generally used for machine translation(MT) to compare the reference sentences to candidate sentences. BLEU scores also utilized to evaluate the poem generation in previous works (Zhang and Lapata, 2014),(Yan, 2016),(Li et al., 2018). We calculated the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores, and we compared the results with the work of (Talafha and Rekabdar, 2019b) and (Talafha and Rekabdar, 2019a). The GPT-2 model better than other models for BLEU-1, BLEU-2, BLEU-3 and BLEU-4 as shown in Table 4.

4.2 Human Evaluation

Given writing poems is a difficult task, there are always inconsistencies between human evaluation and automatic evaluation. We conduct a human evaluation to measure the performance of our model. We invited four experts on Arabic literature to assess the poems generated by our model. We adopted the evaluation in (Zhang and Lapata, 2014), (Li et al., 2018) and (Talafha and Rekabdar, 2019b), and we asked the annotators to evaluate 40 poems generated on four dimensions:

- Fluency (is the generated poem grammatically satisfied?),
- Coherence (Is the generated poem thematically coherent?),
- Meaning (How meaningful the content of a generated poem is?),
- and Poeticness (Does the generated poem have the features of poetry?)

Each dimension is rated on scale 1 (bad) to 5 (excellent). To estimate the annotation reliability, we use Krippendorff's α (Krippendorff, 2013) as Inter-Annotator Agreement (IAA). Krippendorff's α is based on the assumption that expected agreement is calculated by looking at the overall distribution of ratings regardless of the annotator who produced those ratings. Table 7 reports the Krippendorff's α measured for each dimensions. As reported in Table 7, reliabilities ranged from 0.91 and 0.78 indicating the consistency of the annotators ratings.

Table 7: Inter-Annotator Agreement.

Variable	Krippendorff's α
Fluency	0.91
Coherence	0.81
Meaning	0.78
Poeticness	0.83

Table 5 reports the results of human evaluation. We can see that GPT-2 model outperforms the other models in terms of Poeticness and Fluency, and get a good result in terms of Coherence and Meaning compared with other models.

4.3 Generated Examples and Observations

Table 6 shows examples of poems generated by our model. We observe that the model commits to meters and rhyme in many verses. In the sense of following a fixed musical and meter pattern, the model abides by the rhythm controls mainly.

If the previous researches (Talafha and Rekabdar, 2019b; Talafha and Rekabdar, 2019a) on Arabic poem generation produced the best results on the level of meaning and coherence. In these works, the topics were specific: love and religion. Unlike our work, the topics are multiple and comprehensive for most, if not all, the topics of Arabic poetry.

It is also noticeable that the traditional Arabic prevails over the poetry that our model generated, which is a difficult language.

4.4 Ablation Study

We performed ablation on fine-tuning to establish evidence that fine-tuning accounts for a significant boost in performance over the pre-trained model. We used the GPT-2 model we pre-trained on Arabic dataset to generate poems using the same settings as in the prior

Table 8: Ablation BLEU Comparison.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GPT-2 Model + only pre-training	0.5535	0.1737	0.0395	0.0180
GPT-2 Model + pre-training + fine-tuning	0.8739	0.5369	0.3230	0.1871

Table 9: Ablation human evaluation.

Models	Fluency	Coherence	Meaning	Poeticness
GPT-2 Model + only pre-training	1.5	1.3	1.5	0
GPT-2 Model + pre-training + fine-tuning	2.8	2.6	2.6	3.4

Table 10: Example samples generated by GPT-2 + only pre-training.

The generated samples in Arabic	The generated samples in English
حاليا حول موضوع الاتحاد الآسيوي لكرة القدم الذي يشغل في دعم رياضة البولنج المحلية عامة.	Currently on the topic of the Asian Football Confederation who is running in support of the local sport of bowling in general.
هذا العام و قال بوش ان المؤتمر يضم نخبة من المتخصصين في مجال العمل التي يمكن ان تؤدي من خلالها الى تاسيس مثل هذه الدول و من بينهم الرئيس مبارك.	This year, Bush said that the conference includes a group of specialists in the field of work that can lead to the establishment of such countries, including President Mubarak.

experiments. In Tables 8 and 9, we compare the performance of the model with ablation (GPT-2 + only pre-training) to our proposed model (GPT-2 + pre-training + fine-tuning). Table 8 highlights a boost of 58% in BLEU score. In Table 9, we observe that generative model without fine-tuning is noted as zero in terms of poeticness. We also notice that the ablated model observed a significant degradation on fluency, coherence, and meaning. This provides evidence that the fine-tuning step is necessary in achieving state-of-the-art performance. Table 10 shows samples generated by GPT-2 without fine-tuning.

5 CONCLUSION AND FUTURE WORK

This work is the first in the literature to propose pre-training and fine-tuning GPT-2 to automatic Arabic poem generation. In this paper, we use the Khaleej-2004 and Watan-2004 corpora for pre-training the GPT-2 model on Arabic language and use the Arabic poetry dataset in the fine-tuning process to generate Arabic poems. We used the BLEU score to evaluate the performance of poetry generation and human evaluation of four criteria: fluency, coherence, meaning, and poeticness. Both automatic and human evaluations show that our proposed model is good at generating Arabic poetry. The human expert evaluation also demonstrates that our model outperforms baseline models in terms of fluency and poeticness. In the future, we will increase our model performance by focusing on generating specific topics of Arabic poetry.

REFERENCES

Abbas, M. and Smaili, K. (2005). Comparison of topic identification methods for arabic language. In *Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP*, pages 14–17.

Abbas, M., Smaïli, K., and Berkani, D. (2011). Evaluation of topic identification methods on arabic corpora. *J. Digit. Inf. Manag.*, 9(5):185–192.

Bena, B. and Kalita, J. (2020). Introducing aspects of creativity in automatic poetry generation. *arXiv preprint arXiv:2002.02511*.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Das, A. and Gambäck, B. (2014). Poetic machine: Computational creativity for automatic poetry generation in bengali. In *ICCC*, pages 230–238.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Ghazvininejad, M., Shi, X., Choi, Y., and Knight, K. (2016). Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Ghazvininejad, M., Shi, X., Priyadarshi, J., and Knight, K. (2017). Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48.

Guo, G., Ouyang, S., Yuan, F., and Wang, X. (2018). Approximating word ranking and negative sampling for word embedding. In *International Joint Conferences on Artificial Intelligence Organization*.

Hakami, A., Alqarni, R., Almutairi, M., and Alhothali, A. (2021). Arabic poems generation using lstm, markov-lstm and pre-trained gpt-2 models. In *Computer Science & Information Technology (CS & IT)*, volume 11, pages 139–147.

He, J., Zhou, M., and Jiang, L. (2012). Generating chinese classical poems with statistical machine translation models. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1650–1656.

Heilbron, M., Ehinger, B., Hagoort, P., and De Lange, F. P. (2019). Tracking naturalistic linguistic predictions with deep neural language models. *arXiv preprint arXiv:1909.04400*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology (third edition)*. Sage Publications.

Lau, J. H., Cohn, T., Baldwin, T., Brooke, J., and Hammond, A. (2018). Deep-speare: A joint neural model of poetic language, meter and rhyme. *arXiv preprint arXiv:1807.03491*.

Li, J., Song, Y., Zhang, H., Chen, D., Shi, S., Zhao, D., and Yan, R. (2018). Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3890–3900.

Mikolov, T., Karafiát, M., and Burget, L. (2010). Jančernocky, and sanjeev khudanpur. 2010. recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, pages 1045–1048.

Oliveira, H. G. and Cardoso, A. (2015). Poetry generation with poetryme. In *Computational Creativity Research: Towards Creative Machines*, pages 243–266. Springer.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Singh, D., Ackerman, M., and Pérez, R. Y. (2017). A ballad of the mexicas: Automated lyrical narrative writing. In *ICCC*.

Subramanian, S., Rajeswar, S., Dutil, F., Pal, C., and Courville, A. (2017). Adversarial generation of natural language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 241–251.

Talafha, S. and Rekabdar, B. (2019a). Arabic poem generation incorporating deep learning and phonetic cnnsubword embedding models. *International Journal of Robotic Computing*, pages 64–91.

Talafha, S. and Rekabdar, B. (2019b). Arabic poem generation with hierarchical recurrent attentional network. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 316–323. IEEE.

Van de Cruys, T. (2020). Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, Q., Luo, T., and Wang, D. (2016a). Can machine generate traditional chinese poetry? a feigenbaum test. In *International Conference on Brain Inspired Cognitive Systems*, pages 34–46. Springer.

Wang, Q., Luo, T., Wang, D., and Xing, C. (2016b). Chinese song iambics generation with neural attention-based model. *arXiv preprint arXiv:1604.06274*.

Wang, Z., He, W., Wu, H., Wu, H., Li, W., Wang, H., and Chen, E. (2016c). Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*.

Wei, J., Zhou, Q., and Cai, Y. (2018). Poet-based poetry generation: Controlling personal style with recurrent neural networks. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, pages 156–160. IEEE.

Xu, L., Jiang, L., Qin, C., Wang, Z., and Du, D. (2018). How images inspire poems: Generating classical chinese poetry from images with memory networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yan, R. (2016). i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In *IJCAI*, pages 2238–2244.

Yan, R., Jiang, H., Lapata, M., Lin, S.-D., Lv, X., and Li, X. (2013). i, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Yang, X., Lin, X., Suo, S., and Li, M. (2017). Generating thematic chinese poetry using conditional variational autoencoders with hybrid decoders. *arXiv preprint arXiv:1711.07632*.

Yi, X., Li, R., and Sun, M. (2017). Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 211–223. Springer.

Yi, X., Li, R., and Sun, M. (2018). Chinese poetry generation with a salient-clue mechanism. *arXiv preprint arXiv:1809.04313*.

Zhang, J., Feng, Y., Wang, D., Wang, Y., Abel, A., Zhang, S., and Zhang, A. (2017). Flexible and creative chinese poetry generation using neural memory. *arXiv preprint arXiv:1705.03773*.

Zhang, X. and Lapata, M. (2014). Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

Zhou, C.-L., You, W., and Ding, X. (2010). Genetic algorithm and its implementation of automatic generation of chinese songci. *Journal of Software*, 21(3):427–437.

Zugarini, A., Melacci, S., and Maggini, M. (2019). Neural poetry: Learning to generate poems using syllables. In *International Conference on Artificial Neural Networks*, pages 313–325. Springer.