Short Poem Generation (SPG): A Performance Evaluation of Hidden Markov Model based on Readability Index and Turing Test

Ken Jon M. Tarnate[1], May M. Garcia[2], Priscilla Sotelo-Bator[3]
Computer Studies Department, College of Science
Technological University of the Philippines, Manila, Philippines

***Abstract*—We developed a Hidden Markov Model (HMM) that automatically generates short poem. The HMM was trained using the forward-backward algorithm also known as Baum Welch algorithm. The training process was exhausted by a hundreds of iterations through recursion method. Then we used the Viterbi algorithm to decode all the best possible hidden states to predict the next word, and from the previous predicted word, it will generate another word, then another word until it reaches the desire word length that was set in the program. Afterwards, the model was evaluated using several kinds of readability metrics index which measure the reading difficulty and comprehensiveness of the generated poem. Then, we performed a Turing Test, which participated by 75 college students, who are well versed in poetry. They determined if the generated poems was created by a human or a machine. Based from the evaluation results, the highest readability score index of the generated short poem is in the grade 16th level. While 69.2% of the participants in the Turing Test, agreed that most of the machine generated poems were likely created by some well-known poets and writers.**

***Keywords*—Evaluation metrics; Hidden Markov Model; poetry generation; readability test; turing test**

I. Introduction

Hidden Markov Model (HMM) has been successfully explored and applied in the fields of medical technology, military, forensics, bioinformatics, data security and even in arts and literatures. HMM models had also been widely used to create various types of applications software such as speech recognition, image and signal processing, and even text and poetry generation. As of today, Hidden Markov Model (HMM) has now become the based algorithm for creating a text generator, text summarizer, lyrics and music generator [1], [17] These applications is belong into one specific area of natural language processing called computational creativity, where the goal of the artificial intelligence (AI) is to change the nature of creative processes, where the machine will compete with human creativeness in terms of writing through the use of different mathematical models and algorithms.[2],[3] One specific product of this area is called "Poetry Generation." Where the machine will automatically generate poem(s) based from historical data or corpus data used to train the AI model. This complex task required a considerable amount of input knowledge (e.g. phonetics, syntax, semantics, grammar and rhymes). And currently, Hidden Markov Model (hmm) have been successfully conquer this specific topic area of the natural language processing [4], [5]. However, there are still room and subject for improvements, especially on the testing analysis and performance evaluation of the HMM model. Most of the published papers focus on the inner performance of the HMM model and used F1-score, precision and recall to calculate its accuracy. However, only few had considered to measure the outer performance of the HMM model and evaluated the content features of its generated output. In this paper, we tested the learning ability of the hidden markov based on the number of iterations performed before it produces a high quality machine generated poem. [6], [7] Then we evaluated the content of the generated poem by getting the readability score index. And lastly, we performed a standard Turing Test, to confirm the validity and authenticity of the generated poems. Where the participants are asked to determine, whether the generated poem was created by machine or human?

II. Related Works

There is an ample research on the application of Hidden Markov Model in the area of computational creativity and evaluated their model based on the quality and content of the generated composition (e.g. poems, lyrics, short story, novel and etc.) [8], [9].One of these researches is the "Wishful Automatic Spanish Poet" a program created by Pablo Gervás (2000). Where he tested and evaluated his HMM model by calculating the correctness of the generate poem based from the rhymes, syllables and word repetition [8]. Meanwhile, Hugo Oliveira (2008) created a platform for the automatic generation of poetry called "PoeTryMe." He evaluated his model using rhyme and lexical density. Rhyme density was defined as a quantitative measure of the technical quality of the composition. [10] While lexical density is getting the meaningfulness of the composition by identifying the basic parts of speech such as nouns, pronouns, verbs, adverbs, etc. [11], [14] For decades, the literature in poetry generation and the application of Hidden Markov Model in the computational creativity and natural language processing is still uprising and for many years it was already proven that Hidden Markov Model was suitable and efficient to use for text generation. One examples of this are; Polish language text generator [11] Steganographic text based [12] and Chinese couplets generation [13] However, researchers strongly recommended to explore more on the type of testing and what type of other evaluation metrics can be more suitable to evaluate the performance of the HMM model [6], [7]. For this purpose, we explored and used the other evaluation metrics used in computational creativity to evaluate our HMM model.

III. Research Methodology

A. Data Collections and Pre-Processing

A total of 1600 variety poems have collected from different sources; 500 poems extracted manually in the poemhunter.com website, 500 poems that talk about life which come from the poemsforfree.com website, and another 500 poems freeromanticlovepoems.net. We also included the Pablo Neruda collection which contains 100 different love poems. By combining all those poems in a single text file and make it as a corpus dataset. We pre-process the text by omitting unnecessary spaces, symbols and characters such as (&. /, *, “”, (), etc.).

B. Content based Analysis on the Corpus Data

We used text analyzer software, which interpret the content of the collected dataset. Below are the results of the analysis.

C. Training of Hidden Markov Model

We used the forward-backward algorithm also known as Baum Welch algorithm. Then run the program with hundreds of iterations. And using the Viterbi algorithm it decodes the learning and predicts the next words based from the previous generated words. This process was repeated one hundred times to ensure that our model will give a better result compared to the previous results. Afterwards, the program will require an input seed word before it generates another word which will be the basis of the Hidden Markov Model to predict the next word (see Fig. 1 for the details of the model and see Fig. 2 for the actual sample output of the Hidden Markov Model.).

D. Testing of Hidden Markov Model

After the training of the model, we run the program for a hundreds of iterations, until the model produced a better quality of generated poems. In our experiments, we enter a word which are not present on the datasets and tested out if the Hidden Markov Model will still generate text.

Table II shows the numbers of iterations occurred before the Hidden Markov Model (HMM) generates a high quality of short poem. As we observed, the number of iterations is relevant for the generation of a good quality composition. And even though the seed words used in the experiment is not present on the data corpus, the Hidden Markov Model still able to generate short poems.

TABLE. I. Results of Text Analysis

Type of Test	Score
Total Word Count	19842
Number of different words	12727
Complexity factor	27.7%
Readability (Gunning-Fog Index)	7.3
Total number of characters	93521
Number of characters without spaces	52748
Average syllables per word	1.43
Sentence count	11055
Average sentence length	16.56
Max sentence length	188
Min sentence length	1

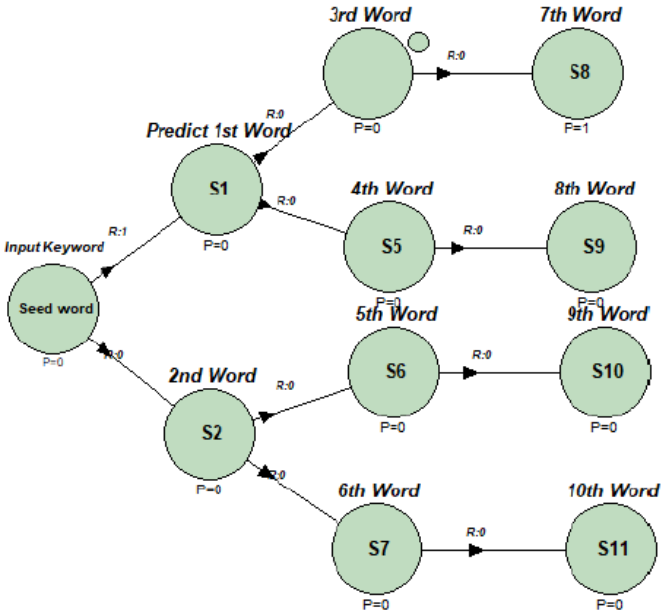

Fig. 1. Design of Hidden Markov Model for Generating Short Poem.

TABLE. II. Testing of Hidden Markov Model using a Seed Word which is not Present on the Corpus Dataset

Seed Word	Generated Poems	No. of Iterations
Education	Education is a room i fear it magnify myself, approaching to forever half of the mind already in my fantasy; it is the truth, that makes me smile, madly but sir, it is the room i fear room, after room, it	Best results after 58 iterations
Money	Money is riches it is up to you to take a rest or fall apart, by lips we call, the person might have joy or fear of it, the love for it is a road to be taken. a thirst from it lies a painful life or maybe a passionate one.	Best results after 110 iterations
Politics	Politics lonely is as love, it must be insist. one is eternal from wisdom trite let your will be different and feels all the gypsy of the men	Best results after 98 iterations
Envy	Envy is not touch not will. i feel them and recall, the truth about liv'd fires the joy of fire it brings harden yet is not touch, to cut is to clean if our hearts make us sea. to keep from pursuing the one i fierce	Best results after 127 iterations
Jealous	Jealous cannot choose you and vague time, and all the breeze, gets gentle the relishing air, fresh of love which rather i would flail, his admittance gained the distractions is more growing until the day	Best results after 128 iterations
Average Iterations		**104.2**

IV. Performance Evaluation

Evaluation of machine generated poem is a vigorously difficult task. Based from the previous researches, most of the machine learning model was evaluated using accuracy metrics such as F1-score, Precision and Recall However, in this study, we used Readability Index to examine the content and comprehensiveness of the generated poem and performed a standard Turing Test to examine the authenticity of the machine generated poem produced by Hidden Markov Model (hmm).

A. Readability Index

Readability Index is defined as the estimate difficulty of a text to read. By measuring the text's complexity by counting the different attributes present on the text such as word lengths, sentence lengths and syllables. There are standards readability metrics used in the United States both in public and private schools and reading centers (e.g. Gunning-Fog, Flesch-Kincaid, Smog, Coleman-Liau and Automated Readability Index) [15], [16].

Based from the results of readability test, the sample generated poem was appropriate and suitable for the Grade 9th to Grade 10th students. This means that the reading difficulty of the entire sample text is in the Highschool level (see Fig. 3). Note: The automated graph in Fig. 3 was generated using text analyzer software powered by "analyzemywriting.com".

Table III shows the results of the readability index. Based from the results the average grade level difficulty of the text is suitable for grade 16th mostly fall for college students and working professionals. This means that the content of the generated poems was really deep and highly constructed.

B. Turing Test

This type of test is defined by Professor Alan Turing in his paper "Computing Machinery and Intelligence" where the AI machine will be tested based on how closely it can resemble or compete to human's intelligent [5], [14]. And finding human evaluators who are technically expert and well versed in writing is a tedious task. For this purpose, we sampled our machine generated poems to the 25 IT students, 25 IS students and 25 Computer Science students of Technological University of the Philippines. All participants were informed that the test was for a research project and instructed them about the Turing Test. We used the Table I sample generated poem as an actual sample to be tested.

As we observed in Table IV, the overall result was positive as the entire generated short poems were able to deceive a significant numbers of participants. Thinking that they were written by human poets, which is the primary goal of this experiment. The generated poem "Money" got the highest positive results. 83% of the participants agreed that this verse was created by some well-known poet or writer. Based from these results, the machine generated poem successfully deceived the human perspective and creativeness where 69.2% of the participants agreed that the poem generated by the Hidden Markov Model can compete to the human creativeness. (see Fig. 4, for the graph results of the conducted Turing Test.)

```
Enter Seed Word: God
Generated Poem:
God best sweetest feels, is whe he gets me
that once again, alone and lost the doors
an open ectasy. to want you now him, for
the schemes, ambitious, underneath darkness  of whole
the only pain and gold is heaven, a brittle self
reveals itself
```

Fig. 2. Sample Generated Poem Tested in a Text Analyzer Software.

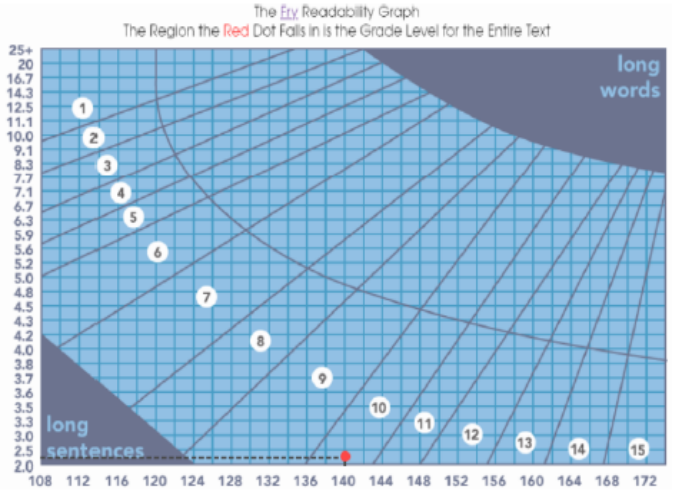

Fig. 3. Results of Readibility Test.

TABLE. III. Results of Readability Test

Type of Test	Grade Level
Gunning-Fog	20.67
Flesch-Kincaid	18.48
SMOG	13.02
Coleman-Liau	9.41
Automated	21.79
Average Grade Level:	**16.68**
Median Grade Level:	**18.48**

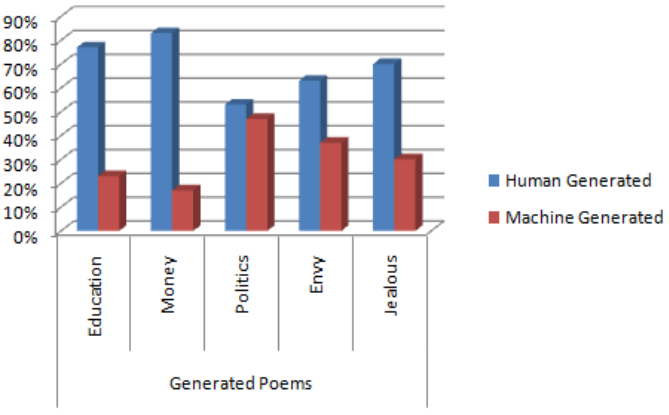

Fig. 4. Results of Turing Test.

TABLE. IV. RESULTS OF TURING TEST (IS THE POEM CREATED BY HUMAN OR BY THE MACHINE?)

Seed Word	Generated Poems	Human Generated	Machine Generated
Education	Education is a room i fear it magnify myself, approaching to forever half of the mind already in my fantasy; it is the truth, that makes me smile, madly but sir, it is the room i fear room, after room, it	77%	23%
Money	Money is riches it is up to you to take a rest or fall apart, and by lips we call, the person might have joy or fear of it, the love for it, is a road to be taken. a thirst from it lies a painful life or maybe a passionate one.	83%	17%
Politics	Politics is as love, its loneliness must be insist. one is eternal from wisdom trite let your will be different and feels all the gypsy of the men	53%	47%
Envy	Envy is not touch not will. i feel them and recall, the truth about liv'd fires the joy of fire it brings harden yet is not touch, to cut is to clean if our hearts make us sea. to keep from pursuing the one i fierce	63%	37%
Jealous	Jealous cannot choose you and vague time, and all the breeze, gets gentle the relishing air, fresh of love which rather i would flail, his admittance gained the distractions is more growing until the day	70%	30%
Average		**69.2%**	**30.8%**

V. CONCLUSIONS

The designed Hidden Markov Model has successfully generated short poems which able to pass the Turing Test and able to deceived the human mind. We able to achieved our objectives of examining the performance of the Hidden Markov Model using only the content features of its generated output and not relying on the accuracy metrics of the model which most of the researchers done. We attack a new type of approach of testing a machine learning model using the environment factors such as human, lexicons and comprehensiveness of the generated text. Currently, the outputs of our model are not yet perfectly correct in terms of grammars and semantics. As for the future works, exploring the semantic and syntactic relations of the words should be a good opportunity to look at, to develop new feature data engineering pipeline and approaches to improve the encoding and decoding process of the Hidden Markov Model.

ACKNOWLEDGMENT

This research was accomplished by the grace and power of our Lord Jesus Christ. All the glory and honor belongs to Him. Special thanks to: Dr. Dionisio A. Espression Jr., Prof. Fidela Q. Aranes, Prof. Fernando L. Renegado and Dr. Ira C. Valenzuela and the entire University Research Development Services (URDS) for making this research possible. – TUP for the World!

REFERENCES

[1] Addanki, K., & Wu, D. (2013, July). Unsupervised rhyme scheme identification in hip hop lyrics using hidden Markov models. In International conference on statistical language and speech processing (pp. 39-50). Springer, Berlin, Heidelberg.

[2] Besold, T. R., Schorlemmer, M., & Smaill, A. (Eds.). (2015). Computational creativity research: towards creative machines.

[3] Petrushin, V. A. (2000). Hidden markov models: Fundamentals and applications. In Online Symposium for Electronics Engineer.

[4] Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. Lund Working Papers in Linguistics, 53, 61-79.

[5] Fernandez, A. C. T., Tarnate, K. J. M., & Devaraj, M. (2018). Deep Rapping: Character Level Neural Models for Automated Rap Lyrics Composition. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2SR. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Kamal, M. S., Chowdhury, L., Khan, M. I., Ashour, A. S., Tavares, J. M. R., & Dey, N. (2017). Hidden Markov model and Chapman Kolmogrov for protein structures prediction from images. Computational biology and chemistry, 68, 231-244.

[7] McCane, B., & Caelli, T. (2004). Diagnostic tools for evaluating and updating hidden Markov models. Pattern recognition, 37(7), 1325-1337.

[8] Gervás, P. (2000, April). Wasp: Evaluation of different strategies for the automatic generation of spanish verse. In Proceedings of the AISB-00 symposium on creative & cultural aspects of AI (pp. 93-100)

[9] Silva, E. D. S., Leão, R. M. M., & Muntz, R. R. (2010, October). Performance evaluation with hidden markov models. In International Workshop on Performance Evaluation of Computer and Communication Systems (pp. 112-128). Springer, Berlin, Heidelberg.

[10] Oliveira, Hugo Gonçalo. "PoeTryMe: a versatile platform for poetry generation." Computational Creativity, Concept Invention, and General Intelligence 1 (2012): 21.

[11] Szymanski, G., & Ciota, Z. (2002). Hidden Markov models suitable for text generation. In WSEAS International Conference on Signal, Speech and Image Processing (WSEAS ICOSSIP 2002) (pp. 3081-3084).

[12] Yang, Z., Jin, S., Huang, Y., Zhang, Y., & Li, H. (2018). Automatically generate Steganographic text based on markov model and Huffman coding. arXiv preprint arXiv:1811.04720.

[13] Pan, Z., Zhang, S., & Guo, Y. (2018). Easycouplet: Automatic generation of Chinese traditional couplets. In Transactions on Edutainment XIV (pp. 117-132). Springer, Berlin, Heidelberg.

[14] Tarnate, K. J. M., & Devaraj, M. (2019) Prediction of ISO 9001: 2015 Audit Reports According to its Major Clauses using Recurrent Neural Networks. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2.

[15] Fernandez, A. C. T. (2019). Computing the Linguistic-Based Cues of Credible and Not Credible News in the Philippines Towards Fake News Detection.

[16] Senter, R. J., & Smith, E. A. (1967). Automated readability index. Cincinnati Univ Oh.

Karaa, W. B. A., & Dey, N. (2017). Mining multimedia documents. Chapman and Hall/CRC.