There Once Was a Really Bad Poet, It Was Automated but You Didn't Know It

Jianyou Wang[1], **Xiaoxuan Zhang**[1], **Yuren Zhou**[2], **Christopher Suh**[1], **Cynthia Rudin**[1,2]

Duke University {[1]Computer Science, [2]Statistics} Department, United States
jw542@duke.edu, zhangxiaoxuanaa@gmail.com
yuren.zhou@duke.edu, csuh09@gmail.com, cynthia@cs.duke.edu

Abstract

Limerick generation exemplifies some of the most difficult challenges faced in poetry generation, as the poems must tell a story in only five lines, with constraints on rhyme, stress, and meter. To address these challenges, we introduce *LimGen*, a novel and fully automated system for limerick generation that outperforms state-of-the-art neural network-based poetry models, as well as prior rule-based poetry models. *LimGen* consists of three important pieces: the Adaptive Multi-Templated Constraint algorithm that constrains our search to the space of realistic poems, the Multi-Templated Beam Search algorithm which searches efficiently through the space, and the probabilistic Storyline algorithm that provides coherent storylines related to a user-provided prompt word. The resulting limericks satisfy poetic constraints and have thematically coherent storylines, which are sometimes even funny (when we are lucky).

1 Introduction

A limerick is a short and catchy 5-line poem that tells a funny, crude, or ironic story. It has strict structural constraints such as an AABBA rhyming scheme, a 99669 syllable count, and an anapestic meter pattern (Legman, 1988). Writing limericks is a challenging task even for human poets, who have to carefully choose, optimize, and even invent new words to satisfy all of the constraints while incorporating creativity and humor.

Prior to this paper, there has not been a successful attempt at realistic automatic limerick generation. Perhaps this is because the task is challenging: Large-scale neural networks often fail to generate decent limericks because the amount of available human-written limericks to learn from is much smaller than other forms of poetry, and because limericks must follow strict structural, meter, and rhyming constraints. Traditional methods for generating limericks instead hard-code the constraints into a template, so that the constraints are obeyed but the generated poems are all extremely similar (resembling Mad Libs, where one fills words into a single template).

In this paper, we introduce a novel system of algorithms for automatic limerick generation, denoted as *LimGen*. *LimGen* takes a user-specified prompt word and produces a creative and diverse set of limericks related to the prompt. Table 1 shows some of *LimGen*'s output.

LimGen is a rule-based search method. Its main components are: (1) Adaptive Multi-Templated Constraints (AMTC), which constrain *LimGen*'s search to a space of realistic limericks, leveraging knowledge from limerick sentence structures extracted from human poets; (2) the novel Multi-Templated Beam Search (MTBS), which searches the space in a way that fosters diversity in generated poems; and (3) the probabilistic Storyline algorithm, which provides coherent storylines that are thematically related to the prompt word.

LimGen relies on the part-of-speech (POS) limerick templates extracted from a small training set and uses a pre-trained language model to fill words into the templates. We used the 345M version of pre-trained GPT-2 (Radford et al., 2019), which performs extremely well in unconstrained text generation. However, it is important to note that a language model such as GPT-2, powerful though it may be, is only a plugin module for *LimGen*. Without *LimGen*, GPT-2 alone is completely incapable of generating limericks.

Through our experiments, we demonstrate that *LimGen* creates a new benchmark for limerick generation, outperforming both traditional rule-based algorithms and encoder-decoder style neural networks across a variety of metrics, including

Transactions of the Association for Computational Linguistics, vol. 9, pp. 605–620, 2021. https://doi.org/10.1162/tacl_a_00387
Action Editor: Micha Elsner. Submission batch: 10/2020; Revision batch: 1/2021; Published 7/2021.

(**a**) prompt: ‘‘money’’
There was a greedy man named Todd, Who lost all his money in a fraud. When he returned to work, He was robbed by his clerk, And never could buy a cod.
(**b**) prompt: ‘‘cunning’’
There was a magician named Nick, Who fooled all his family in a trick. When he returned to hide, He was found by his bride, And killed with a magical lipstick.

Table 1: *LimGen* examples.

emotional content, grammar, humor, sensibleness and storyline quality. Furthermore, although *LimGen* is not yet on par with human poets, our experiments show that 43% of *LimGen*'s output cannot be distinguished from human-written limericks even when directly compared with actual human limericks.

The main contributions of this paper are the multi-template-guided *LimGen* system and its MTBS search algorithm. Equipped with AMTC, *LimGen* is the first fully automated limerick generation system that has the ability to write creative and diverse limericks, outperforming existing state-of-the-art methods. Our diversity-fostering beam search (MTBS) is on par with some of the best beam search algorithms in terms of its ability to optimize limerick quality, and it does a significantly better job at fostering diversity than other methods. The code for *LimGen* as well as the complete list of machine-generated limericks used in our experiments are available online (Wang et al., 2020).

From a broader perspective, we have shown that rule-based poetry generation systems that follow a multi-templated approach, as implemented via the AMTC in this work, can perform better than large-scale neural network systems, particularly when the available training data are scarce. Our work indicates that a computational system can exhibit (what appears to be) creativity using domain-specific knowledge learned from limited samples (in our case, POS templates extracted from human-written poems). Although we only use templates to capture the part-of-speech structure of limericks, in general, templates can represent any explicit or latent structures that we wish to leverage. Other NLP applications (e.g., biography generation, machine translation, image captioning, machine translation) have also seen revived interest in template-guided approaches (Wiseman et al., 2018; Yang et al., 2020; Deshpande et al., 2019; Wang et al., 2019). Thus, it is conceivable that the general framework of *LimGen*, including AMTC and the MTBS algorithm, can be applied to other forms of poetry generation, as well as broader domains in NLP.

2 Related Literature

To the best of our knowledge, Poevolve (Levy, 2001), which combines an RNN with an evolutionary algorithm, and the stochastic hill climbing algorithm proposed by Manurung et al. (2000) are the only other serious attempts at limerick generation in the past 20 years. Unfortunately, their implementations did not result in readable limericks, as can be seen in Section 4.3.

Traditional rule-based methods in poetry generation are able to enforce hard constraints, such as rhyming dictionaries or POS templates (Gervás, 2000, 2001; Colton et al., 2012; Yan et al., 2016). *LimGen* is also rule-based, though it has substantially more flexibility and diversity than Colton et al.'s (2012) approach, which follows a single POS template during poetry generation. Needless to say, the use of adaptive multi-templates makes AMTC the bedrock of *LimGen*.

Neural language models have recently been able to produce free-style (unconstrained) English poetry with moderate success (Hopkins and Kiela, 2017; Liu et al., 2018). In Chinese poetry generation (Zhang and Lapata, 2014; Yi et al., 2018b; Wang et al., 2016; Yi et al., 2018a), research has been so successful that it has spurred further efforts in related areas such as sentiment and style-controllable Chinese quatrain generation (Yi et al., 2020; Yang et al., 2018; Chen et al., 2019). However, their large-scale neural network models take advantage of the Chinese quatrain database, which has more than 150k training examples. In contrast, *LimGen* uses less than 300 limericks. Most modern poetry-generation systems are encoder-decoder style recurrent networks (e.g., character-level and word-level LSTMs) with modifications such as various forms of attention mechanisms. Lau et al. (2018) integrated these techniques and proposed *Deep-speare*, which represents the state-of-the-art for Shakespearean sonnet generation. In our

experiments, we have adapted and re-trained *Deep-speare* for limerick generation. Empirically, it cannot compete with *LimGen*.

For handling rhyming constraints, unlike Ghazvininejad et al. (2016) and Benhart et al. (2018), who generate the last word of each line before generating the rest of the line, our proposed Storyline algorithm selects a probability distribution for the last word of each line.

Beyond poetry generation, templates are often used in other NLP tasks. For biography generation, Wiseman et al. (2018) noted that a template-guided approach is more interpretable and controllable. Yang et al. (2020) stated that templates are beneficial for guiding text translation. For fostering diversity in generated text, Deshpande et al. (2019) found that a part-of-speech template-guided approach is faster and can generate more diverse outputs than the non-templated diverse beam search of Vijayakumar et al. (2018). *LimGen*'s MTBS generates diverse results by design; it also addresses the problem of degradation of performance when beam size grows larger, which has been a challenge noted in several prior works (Cohen and Beck, 2019; Vinyals et al., 2016; Koehn and Knowles, 2017).

Since all rule-based constraints in *LimGen* are easily enforced by a filtering function, it does not need to borrow any advanced techniques from the area of constrained text generation (e.g., Hokamp and Liu, 2017; Anderson et al., 2017; Post and Vilar, 2018; Yan et al., 2016) where constraints are more complicated.

3 Methodology

We first introduce terminology in Section 3.1. We present *LimGen* along with AMTC in Section 3.2. We present the MTBS algorithm in Section 3.3, and present our Storyline algorithm in Section 3.4.

3.1 Terminology

We first introduce some useful notation for the concepts of (partial) line, (partial) template, language model, filtering function, and scoring function.

LimGen's entire vocabulary is $\mathbb{W}$ with size $|\mathbb{W}|$. For word $w \in \mathbb{W}$, its POS is $w.pos$. The first t words of a complete line s_i forms a *partial line* $s_i^{(t)}$. We store many partial lines $s_i^{(t)}$ with length t in a set $S^{(t)} = \{s_i^{(t)}\}_i$. A new word w concatenated to $s_i^{(t)}$ becomes $s_i^{(t+1)} = (s_i^{(t)}, w)$. A (partial) template is a sequence of POS tags. The (partial) template of line s is $s.pos = (w_1.pos, \ldots, w_n.pos)$. A language model $\mathcal{L}$ processes a (partial) line and gives a probability distribution for the next word. We use $\mathcal{D}^{(t)}$ to denote the probability distribution at step t. The filtering function $\mathcal{F}$ filters out words that do not satisfy meter and stress constraints by setting the probability mass of these words in $\mathcal{D}^{(t)}$ to zero. Since limericks have a specific meter and stress pattern, words that break this pattern are filtered out by $\mathcal{F}$.

The scoring function for lines is denoted $H(\cdot)$, which is the average negative log likelihood given by the language model. Although our search algorithm generally aims to maximize $H(\cdot)$, the language model's scoring mechanism may not be aligned with poetic quality; sometimes a slightly lower scoring poem has better poetic qualities than a higher scoring one. Thus we may find a better poem by sifting through *LimGen's* output, rather than choosing the highest scoring poem.

3.2 Adaptive Multi-Templated Constraint (AMTC)

Because the number of limericks that exist in available databases is so limited, we cannot expect that a neural network would learn the POS constraints for a valid limerick. Instead, we use rule-based POS constraints, which are useful in that they ensure the output adheres to the known structure of poetry. The use of adaptive multi-templates makes the poems more diverse and interesting by providing *LimGen* with greater flexibility in pursuing many different templates.

It may seem natural to choose multiple templates and have the limerick generation process follow each one of them in parallel, but this is inefficient; instead, we start generating from one template, keeping also the set of templates that agree with what we have generated so far. This way, we generate each line by combining a set of related templates. Specifically, AMTC constrains *LimGen* to consider word w for partial line $s^{(t)}$ only if the template of $s^{(t+1)} = (s^{(t)}, w)$ matches with a human-written template up to the first $t+1$ tokens. Therefore, the more templates we extract from real poems, the higher the degree of freedom we offer *LimGen*.

Algorithm 3.1 *LimGen* with AMTC

▷ In Section 3.4 the Storyline Algorithm describes how storylines are integrated with *LimGen* to generate last words of each line.
Initialize $S^{(0)} \leftarrow \{[\,]\}$;
for $t = 0, 1, 2, \ldots$ **do**
  ▷ $\tilde{S}^{(t+1)}$ will store all candidate partial lines of length $t+1$
  ▷ $S^{(t+1)}$ will store the chosen partial lines by MTBS
  Initialize $S^{(t+1)} \leftarrow \varnothing, \tilde{S}^{(t+1)} \leftarrow \varnothing$;
  for $s_i^{(t)} \in S^{(t)}$ **do**
    ▷ Filter the distribution $\mathcal{L}(s_i^{(t)})$ given by GPT-2
    ▷ for meter and stress match
    $\mathcal{D}_i^{(t+1)} \leftarrow \mathcal{F}(\mathcal{L}(s_i^{(t)}))$;
    for $w_k \in \mathbb{W}$ with $\mathcal{D}_i^{(t+1)}(w_k) > 0$ **do**
      ▷ AMTC ensures partial template is always the
      ▷ prefix of a viable human template
      If $[s_i^{(t)}, w_k]$ satisfies AMTC:
      ▷ if w_k is the last word (i.e., $[s_i^{(t)}, w_k]$'s template
      ▷ matches a human template), where Storyline
      ▷ Algorithm contributes to generation of w_k
      Concat $[s_i^{(t)}, w_k]$, union with $\tilde{S}^{(t+1)}$;
    end for
  end for
  ▷ Find top N lines using multi-templated beam search
  $\tilde{s}_1^{(t+1)}, \ldots, \tilde{s}_N^{t+1} \leftarrow$ *MTBS*$(\tilde{S}^{(t+1)})$;
  Union with $S^{(t+1)}$;
end for
Output S^{t+1}

We present the entire *LimGen* system with AMTC in Algorithm 3.1.

We illustrate *LimGen* with the example in Figure 1. At the 3rd step of generating the second line, set $S^{(3)}$ contains partial lines $s_1^{(3)}$ =*''who ate a''* and $s_2^{(3)}$ =*''who bought a''*, which share the same partial template *''WHO VBD A''*. The probability distributions for the fourth words are $\mathcal{D}_1^{(4)} = \mathcal{L}(s_1^{(3)})$ and $\mathcal{D}_2^{(4)} = \mathcal{L}(s_2^{(3)})$. One can see how *LimGen* with AMTC does not follow a single template using the example in Figure 1 since the partial template *''WHO VBD A''* branches into two distinct partial templates *''WHO VBD A JJ''* and *''WHO VBD A NN''*.

After $\mathcal{F}$ filters out all unsatisfactory words that break the syllable or stress pattern, we obtain two filtered distributions $\tilde{\mathcal{D}}_1^{(4)} = \mathcal{F}(\mathcal{D}_1^{(4)})$ and $\tilde{\mathcal{D}}_2^{(4)} = \mathcal{F}(\mathcal{D}_2^{(4)})$. We then filter out words that do not satisfy the AMTC. The concatenation step in *LimGen* saves all possible partial lines into a temporary set $\tilde{S}^{(4)}$.

The MTBS algorithm then finds N diverse and high-scoring candidate lines from $\tilde{S}^{(4)}$ and saves them into $S^{(4)}$. In the next section, we present the MTBS algorithm in detail.

3.3 Multi-Templated Beam Search (MTBS)

At iteration t, suppose we have a set of partial lines $S^{(t)}$ with size N. Let $\tilde{S}^{(t+1)}$ be the set of all possible one-word extensions of these partial lines. Given the scoring function $H(\cdot)$, a standard beam search would sort $\tilde{S}^{(t+1)}$ in descending order and keep the top N elements. In limerick generation using standard beam search, we also observed the phenomenon documented by Li and Jurafsky (2016) that most of the completed lines come from a single highly-valued partial line. As mentioned before, the innovation of MTBS over previous diverse beam search papers (Li and Jurafsky, 2016; Vijayakumar et al., 2018) is that it calculates a diversity score between (partial) templates (runtime $O(N^2)$), which is more computationally efficient than an approach that assigns a notion of diversity between individual lines (runtime $O(N|\mathbb{W}|)$, $N \ll |\mathbb{W}|$, where N is the total number of templates and $|\mathbb{W}|$ is the vocabulary size). Our proposed diversity score also more accurately captures the diversity between generated lines. The intuition is that if two generated lines have very different templates, they are usually fundamentally different in terms of the progression of the story.

We use a weighted hamming distance to measure the difference between (partial) templates of the same length, denoted as ''diversity score.'' Before formally defining diversity score, we calculate the weights of each POS category. For each POS category, we take the inverse of its percentage of occurrence within all those n^{th} line templates (e.g., second line templates) extracted from the n^{th} line of one of our human-written limericks from the database. (The POS weights are only calculated once, before we start generating the n^{th} line.) We then use the softmax to transform them into weights for each POS, which measure how rare these POS categories are. The softmax nonlinear transformation softly clips the large weights of outlier POS categories that appear only once or twice. More formally we have:

Definition 1 (**Part-of-Speech Weight**). *Let P be the set of all POS categories that occur in all the n^{th} line complete-line templates extracted from the limerick database. $|P|$ is its size. For $p_i \in P$, the proportion of p_i is $q_i = \frac{\#p_i \text{ occurrences}}{\sum_{p_j \in P} \#p_j \text{ occurrences}}$, and the weights of $\{p_i\}_{1 \le i \le |P|}$ are defined as*

$$\{w(p_i)\}_{1 \le i \le |P|} = \textit{softmax}\Big(\{1/q_i\}_{1 \le i \le |P|}\Big).$$

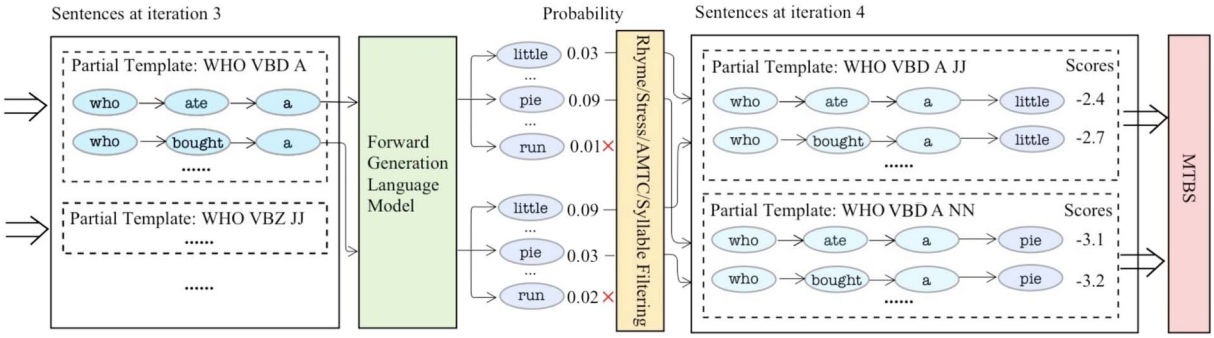

Figure 1: General framework of *LimGen*.

Algorithm 3.2 Multi-Templated Beam Search (MTBS)

▷ The following describes MTBS at iteration t
Input $\tilde{S}^{(t+1)}$
Initialize $S^{(t+1)} \leftarrow \emptyset$, $A \leftarrow \emptyset$;
▷ A will hold the templates we have chosen
Split $\tilde{S}^{(t+1)}$ by templates into m subsets: $\{T_1 : \tilde{S}_1^{(t+1)}, \ldots, T_m : \tilde{S}_m^{(t+1)}\}$,
▷ Lines in the same subset share the same template
For each $\tilde{S}_i^{(t+1)}$, we calculate score h_i by averaging its top n lines according to $H(\cdot)$
$B \leftarrow \{T_1 : h_1, \ldots, T_i : h_i, \ldots T_m : h_m\}$;
▷ In B, each template corresponds to an aggregate score h
Assume $h_j = \max B$, append top n lines according to $H(\cdot)$ from $\tilde{S}_j^{(t+1)}$ to $S^{(t+1)}$;
Delete h_j from B, append T_j into A;
while $|S^{(t+1)}| \leq N - n$ and $B \neq \emptyset$ **do**
  $x \in \text{argmax}_i \left(h_i \sum_{T_k \in A} \|T_i - T_k\|_{div} \} \right)$;
  Append top n lines from $\tilde{S}_x^{(t+1)}$ to $S^{(t+1)}$;
  Delete h_x from B, append T_x into A;
end while
Return $S^{(t+1)}$;

Definition 2 (Diversity Score). *For (partial) templates* $T_1 = \{pos_{11}, \ldots, pos_{1n}\}$ *and* $T_2 = \{pos_{21}, \ldots, pos_{2n}\}$*, assume index set* $A = \{i | pos_{1i} \neq pos_{2i}\}$*, then we define the diversity score (weighted hamming distance) between* T_1 *and* T_2 *as*

$$\|T_1 - T_2\|_{div} = \sum_{i \in A} \max(w(pos_{1i}), w(pos_{2i})).$$

Consider a scenario where (partial) templates T_1 and T_2 have different POS categories at index i but both categories are fairly common (for instance, one noun and one verb), and where (partial) templates T_1 and T_3 also have different POS categories at index j but one or both are rare. Our proposed diversity score will ensure that the diversity between T_1 and T_3 is greater than the diversity between T_1 and T_2, which aligns with our intuition.

In short, given the set of partial lines $\tilde{S}^{(t+1)}$, MTBS will choose N lines, denoted as $S^{(t+1)}$, such that they are high-scoring and generated using many diverse templates. Specifically, we divide $\tilde{S}^{(t+1)}$ into m subsets $\{T_1 : \tilde{S}_1^{(t+1)}, \ldots, T_m : \tilde{S}_m^{(t+1)}\}$, where each subset corresponds to a unique (partial) template of length $t+1$. According to scoring function $H(\cdot)$, for each of these subsets $\tilde{S}_i^{(t+1)}$, we calculate its aggregate score h_i by averaging its n highest-scoring lines. For ease of notation, we let $B = \{T_1 : h_1, \ldots, T_i : h_i, \ldots T_m : h_m\}$, and we initialize $A = \emptyset$ to be the set of previously chosen templates. At this point, we shall iteratively determine the order by which lines from these m subsets will be included into $S^{(t+1)}$.

We select the first subset that has the highest aggregate score within $\{h_1, \ldots, h_m\}$. Assume it is $\tilde{S}_j^{(t+1)}$ with score h_j. We then delete $T_j : h_j$ from B, add T_j to A, and add the top n lines from $\tilde{S}_j^{(t+1)}$ to $S^{(t+1)}$. Then, for each iteration > 1, we calculate a set of temporary new scores $\tilde{B} = \{T_1 : \tilde{h_1}, \ldots, T_i : \tilde{h_i}, \ldots T_m : \tilde{h_m}\}$ where each $\tilde{h_i}$ is the original score h_i multiplied by $\sum_{T_k \in A} \|T_i - T_k\|_{div}$, which is the sum of the diversity scores between T_i and all previously chosen templates in A. These scores are designed to strike a balance between finding high probability lines (as approximated by h) and lines whose templates have high diversity from the previously chosen templates (as measured by $\sum_{T_k \in A} \|T_i - T_k\|_{div}$). Afterwards, we repeat the process of choosing the template with the highest $\tilde{h}$ score, delete it from B, add it to A, and add the top n lines from its corresponding subset to $S^{(t+1)}$. We stop the iteration before the size of $S^{(t+1)}$ exceeds N.

Empirically, MTBS does not favor the templates with the rarest POS (largest distance from the rest), since those will have very low scores from $H(\cdot)$. It turns out MTBS picks templates that are reasonably different from each other while ensuring their generated lines have enough high scores.

3.4 Storyline Algorithm

We define the storyline of a limerick to be the last words of each line, denoted as $Y = (y_1, y_2, y_3, y_4, y_5)$, where y_1 is traditionally a name or place. In addition to making sure Y has an ''AABBA'' rhyming pattern, our storyline algorithm also helps *LimGen* to maintain a consistent theme throughout its process of limerick generation.

We define the probabilistic distribution of storyline Y given a prompt word y_0 as:

$$\begin{aligned} p(Y|y_0) = {} & p(y_2|y_0)p(y_3|y_0)p(y_4|y_0,y_2,y_3) \\ & \cdot p(y_5|y_0,y_2,y_3)p(y_1|y_5), \end{aligned} \tag{1}$$

$$\begin{aligned} p(y_2|y_0) &\propto \mathsf{Sim}(y_2, y_0), \\ p(y_3|y_0) &\propto \mathsf{Sim}(y_3, y_0), \\ p(y_4|y_0,y_2,y_3) &\propto \mathbf{1}^{(r)}_{y_4,y_3} \sum_{i\in\{0,2,3\}} \mathsf{Sim}(y_4, y_i), \\ p(y_5|y_0,y_2,y_3) &\propto \mathbf{1}^{(r)}_{y_5,y_2} \sum_{i\in\{0,2,3\}} \mathsf{Sim}(y_5, y_i), \\ p(y_1|y_5) &\propto \mathbf{1}^{(r)}_{y_1,y_5} \cdot \mathbf{1}^{(p)}_{y_1}. \end{aligned} \tag{2}$$

where the conditional distribution of each storyline word y_i is a multinomial distribution over $\mathbb{W}$. $\mathsf{Sim}(w_1, w_2)$ calculates the semantic similarity between words $w_1, w_2 \in \mathbb{W}$, which is their distance in a pretrained word embedding space. Indicator function $\mathbf{1}^{(r)}_{w_1,w_2}$ denotes whether w_1 rhymes with w_2 and $\mathbf{1}^{(p)}_{w_1}$ denotes whether w_1 is a person's name. By sampling the storyline from $p(Y|y_0)$, we guarantee the following:

- y_2 and y_3 are semantically related to y_0;
- y_4 rhymes with y_3; y_5, y_1 and y_2 rhyme;
- y_4, y_5 are semantically related to y_0, y_2, y_3.

Examples of samples from Storyline's distribution are provided in Table 2.

During the process of generating a limerick, the Storyline algorithm will sequentially generate many storylines in the order of y_2, y_3, y_4, y_5, y_1, each of which satisfies not only the rhyming constraint but also the constraints on POS template, syllable count, and anapestic meter pattern. Figure 2 shows the directed acyclic graph for the Storyline algorithm when the beam size of MTBS is 1.

Prompt	Storyline
war	(Wade, raid, campaign, again, stayed)
sports	(Pete, street, school, pool, athlete)
monster	(Skye, guy, scary, carry, pie)
forest	(Shea, day, friend, end, way)

Table 2: Examples of storyline samples.

In general, given a prompt word y_0, we start by generating the first line l_1, which has a canonical form, with a random name filled at its end as a placeholder for y_1 (otherwise, a pre-specified name will limit the options for y_2, y_5). Following l_1, we use *LimGen* to generate a set of second lines $\{\dots, l_2, \dots\}$ (as described in Algorithm 3.1) with last word left blank. We use $l'_{1:2}$ to denote a limerick generated up to this point (i.e., l_1 concatenated with an almost-finished l_2). For each $l'_{1:2}$, we repeatedly sample y_2 from the conditional Storyline distribution $p(y_2|y_0)$ in (2) until it satisfies constraints on POS, syllable and meter. $[l'_{1:2}, y_2]$ together form the complete first two lines of a limerick. Continuing to generate lines, we use MTBS (described in Algorithm 3.2) to maintain a set of high-scoring and diverse second lines $\{\dots, [l'_{1:2}, y_2], \dots\}$. Note that our language model also assigns a probability score for each y_2. We can continue generating $l'_{1:k}$ with *LimGen* and sampling y_k from the conditional Storyline distribution for $k = 3, 4, 5$ in a similar fashion. Finally, we sample y_1 from $p(y_1|y_5)$ and replace the random name at the end of l_1 by it. The result is a set of limericks $\{\dots, L, \dots\}$, from which we choose the highest scoring one.

4 Experiment

4.1 Experimental Setup

To implement *LimGen*, a significant amount of effort has gone into adapting existing NLP technologies for limerick generation. In order to extract POS templates from human written limericks, we modified the POS categories in NLTK (Bird et al., 2009) by refining certain categories for better quality in our generated limericks. Leveraging NLTK's POS tagging technique, we obtained a list of refined POS templates from a

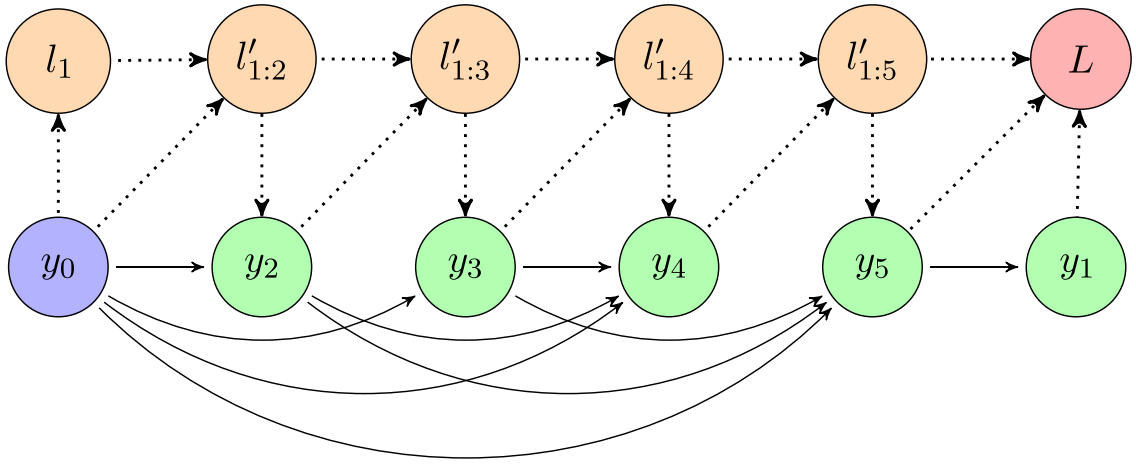

Figure 2: Directed acyclic graph for generating a limerick L with Storyline algorithm given prompt y_0, with solid arrows representing dependency through Storyline distribution (1), shaded arrows representing the generation process of *LimGen*, and the total beam size of MTBS set to be 1 for simplicity.

small limerick dataset of 200 human-written limericks from Cicchi (2019) and Limericks (2019). For a full list of our modified POS categories see Wang et al. (2020). Since the filtering function $\mathcal{F}$ requires knowing each word's syllable and stress pattern, we use CMU (2019) for information on syllable and stress.

As for the implementation of the Storyline algorithm, there are several existing technologies to indicate whether two words rhyme with each other. For example, *Deep-speare* (Lau et al., 2018) proposed a character-level LSTM to learn rhyming. For the sake of simplicity and accuracy, we used a rhyming dictionary curated from Beeferman (2016). We also used a dictionary of names (Namepedia, 2019) from which the Storyline algorithm can choose y_1, the name in the poem's first line. To calculate the semantic similarity between two words, we use the pre-trained word embedding space from *spaCy*'s model (Honnibal et al., 2020).

Note that Algorithm 3.1 is only responsible for generating the last four lines of a limerick. Since first lines of limerick usually have a canonical form, we generate the first lines separately.

The outline of this section is as follows. We first show why GPT-2—or even retrained GPT-2—cannot produce limericks without *LimGen*. We then show the low-quality output from prior attempts at limerick generation. We have also designed five experiments to compare the quality of *LimGen*'s output with limericks from human poets, baseline algorithms, and other state-of-the-art poetry systems re-purposed for limerick generation. All five experiments were evaluated on Amazon Mechanical Turk by crowd-workers, following a protocol similar to that of Lau et al. (2018) and Hopkins and Kiela (2017) (see Section 4.4 for details). Additionally, an ''Expert Judgment'' experiment was conducted where more experienced judges directly evaluated the performance of *LimGen*'s output and human-written limericks across a variety of metrics (See Section 4.8 for details).

Since *LimGen* has three major components: AMTC, MTBS, and Storyline, we designed three baseline algorithms for an ablation analysis in order to investigate the effectiveness of each of them.

-*Single-Template*: MTBS+Storyline but without AMTC

-*No-Story*: AMTC+MTBS but without pre-selected storylines

-*Candidate-Rank*: AMTC+Storyline but we have replaced the MTBS algorithm with another modified beam search algorithm Candidate-Rank (Cohen and Beck, 2019).

In our experiments, *LimGen* and all baseline algorithms use a total beam size of $N = 360$ at each step, MTBS algorithm's individual beam size per template is $n = 12$, and we take the highest scoring poem from the set of output poems. For implementation details please refer to our online GitHub repository (Wang et al., 2020).

4.2 GPT-2 Cannot Generate Poems by Itself

A naïve implementation of GPT-2 simply cannot produce original and valid limericks. GPT-2 tends

(**a**) Two examples of Naïve GPT-2
There was a kind girl whose name is Jane, A girl who I did not know, He then added, She had tons of luggage, It seemed I could walk where she.
(**b**) This output is an exact replica of a human limerick (Vaughn, 1904) in the training corpus of GPT-2.
Wait, there was a young lady in china, Who was quite a greedy young diner. She feasted on snails, Slugs, peacocks and quails, 'No mixture,' she said, 'could be finer.'

Table 3: Two examples of Naïve GPT-2.

to generate long sentences that exceed the syllable limit for limericks. To meet a syllable constraint, we would need to truncate the generated sentences, which creates lines that do not end correctly. Rhyming is insurmountable if we do not utilize additional algorithms, as evidenced by Example (a) of Table 3. The output lacks poetic quality since the training corpus of GPT-2 does not mainly consist of limericks or other kinds of poetry.

If we try to re-train the last few layers of a GPT-2 model on our entire limerick dataset, it does not solve the problem. To our knowledge, our entire dataset is the largest and most comprehensive limerick dataset, consisting of more than 2000 limericks from several sources (Cicchi 2019; Limericks, 2019; Lear, 2010; Parrott, 1984; Haynes, 2010). Even though this dataset is much larger than the subset of data (≈ 300) from which we extracted templates, it is still insufficient to retrain GPT-2. The result of re-training is that GPT-2 severely overfits. It only regurgitates limericks from the training corpus, as seen in Example (b) of Table 3.

Terminating training early (in order to avoid memorization or overfitting) leads only to an awkward merge of problems shown in the two examples of Figure 3 in which the model has not learned enough to faithfully reproduce the form of a limerick, but also often loses coherence abruptly or regurgitates the training set.

Just as *LimGen* needs a powerful pre-trained language model such as GPT-2, without *LimGen*'s algorithms, GPT-2 by itself is unable to accommodate the constraints of limerick generation due to the deficiency of training data.

4.3 Prior Attempts at Limerick Generation

Levy (2001) stated that ''the current system produces limericks that in many ways seem random.'' We have re-run their implementation, and it only produced meaningless verses with serious grammatical issues. Manurung et al. (2000) stated that their work is unfinished and stated that their results ''can hardly be called poems'' (see examples in Table 4). Empirically, *LimGen* has a clear advantage over both prior works. Therefore, the low-quality output from these system do not warrant an extensive comparison with *LimGen*'s poems.

On the other hand, popular internet poem generators (PoemGenerator, 2019; PoemOfQuotes, 2019) have a set of human-written half-finished sentences that are assembled with user input words to create limericks (see Table 5). However, because so little of the resulting limerick is generated by a machine, we cannot consider these internet poem generators as automatic limerick generation systems.

4.4 Experiment 1: *LimGen* vs. *No-Story*

As we have mentioned before, the *No-Story* baseline still utilizes the AMTC and MTBS algorithms. This experiment demonstrates the importance of having pre-selected storylines in poetry generation.

We randomly selected 50 pairs of limericks, in which each pair of limericks consists of one generated by *LimGen* and another generated by *No-Story* using the same prompt word. For each pair of limericks, 5 different crowd-workers (each with an approval rate $\geq 90\%$) answered a list of 6 questions on different evaluation metrics (humor, sensibleness, storytelling, emotional content, grammar, thematic relatedness to prompt) and an additional sanity-check question to filter out illogical responses. Figures 3 and 4 show the side-by-side comparison of a pair of limericks and the list of questions exactly as they appeared on the Amazon Mechanical Turk survey. Note that the output of *LimGen* has a 50% chance of being either Limerick A or B to avoid any left–right bias.

A total of 250 response were recorded, and a small number of responses were filtered out since they did not answer the sanity check question correctly, which asks crowd-workers to count the number of 3-letter words in the fourth line of Limerick B. We have transformed the response such that a response of 5 always means that the poem is

Figure 3: Side-by-side comparison of two limericks generated from different methods.

(**a**) Example of Levy's (2001) system
Ears christmas new throat boat apparel, Plain always obsessed deal idea, Attempt blast work many, Mercator aghast, Kathleen mind revealed barge bugs humor.
(**b**) Example of Manurung et al.'s (2000) system
The bottle was loved by Luke. a bottle was loved by a dog A warm distinctive season humble mellow, smiled refreshingly slowly. Ran.

Table 4: Two prior attempts at limerick generation.

(**a**) Example of PoemGenerator (2019)
There once was a **man** called **Liam**. He said, "See the coliseum!", It was rather young, But not very zedong, He couldn't resist the mit im.
(**b**) Example of PoemOfQuotes (2019)
There was a **man** from White Who liked to **fly his kite** On each sunny day The man would say 'Oh, how I miss White!'

Table 5: Examples of internet poem generators. Underlined parts are human-written half sentences and bold parts are user inputs.

rated as ''Definitely *LimGen*'s output;'' that is, if *LimGen* produced Limerick B, we transform 5 to 1, 4 to 2, 2 to 4 and 1 to 5. After this transformation, we calculated the mean and standard deviation for each metric. Since all questions ask crowd-workers to compare the qualities of two limericks, the results are relative. It should be clear that for any metric, an average greater 3 means *LimGen* is performing better than the baseline method on that metric. To be precise, if the mean of a metric

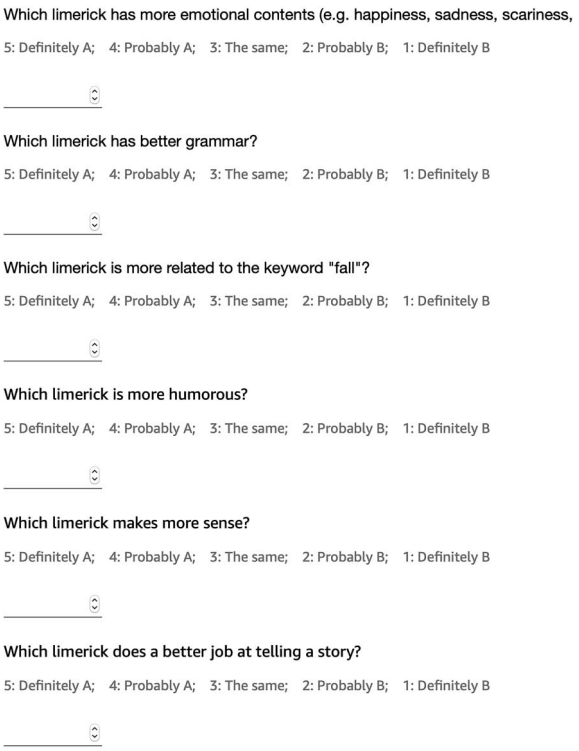

Figure 4: The list of questions on Amazon Mechanical Turk survey.

is > 3, we run a one-sided t-test with the null-hypothesis being ''metric ≤ 3, i.e., *LimGen* is not doing better than baseline.'' If the mean of a metric is < 3, suggesting the baseline is probably doing better, we run the complementary one-sided t-test with the null-hypothesis being ''metric ≥ 3, i.e., baseline is not doing better than *LimGen*.''

From Table 6, the p-value of grammar, humor, relatedness to prompt, and storytelling are all small enough to reject the null hypothesis, which shows that *LimGen* was better than *No-Story* in all four categories. We can weakly reject the null hypothesis for the sensibleness metric, which shows that *LimGen* also may outperform *No-Story* with respect to sensibleness. However, the p-value of emotion is 0.38, therefore we cannot

Metrics \ Statistics	mean	sd	p-value
emotion	3.03	1.22	0.38
grammar	**3.18**	1.27	0.03
humor	**3.14**	1.20	0.05
relatedness	**3.32**	1.22	2.0×10^{-4}
storytelling	**3.35**	1.38	3.0×10^{-4}
sensibleness	3.14	1.42	0.09

Table 6: *LimGen* vs. *No-Story*.

Metrics \ Statistics	mean	sd	p-value
emotion	**3.20**	1.23	0.02
grammar	**3.48**	1.28	4.4×10^{-6}
humor	**3.27**	1.16	0.001
relatedness	2.95	1.42	0.34
storytelling	**3.52**	1.39	1.3×10^{-6}
sensibleness	**3.40**	1.36	9.8×10^{-5}

Table 7: *LimGen* vs. *Single-Template*.

Metrics \ Statistics	mean	sd	p-value
emotion	2.88	1.20	0.62
grammar	3.06	1.14	0.25
humor	2.91	1.15	0.15
relatedness	3.03	1.22	0.37
storytelling	3.06	1.31	0.28
sensibleness	**3.19**	1.27	0.034

Table 8: *LimGen* vs. *Candidate-Rank*.

claim *LimGen*'s output has better emotional content than *No-Story*. Overall, we see that *LimGen* empirically outperforms *No-Story* in 5 categories. From this experiment, we see that having preselected storylines not only makes the limerick more related to the prompt (as expected), but it also enhances the consistency of story-telling and other important poetic metrics.

All other experiments were designed in the same way as Experiment 1.

4.5 Experiment 2: *LimGen* vs. *Single-Template*

As we have mentioned before, the *Single-Template* baseline still utilizes the MTBS and Storyline algorithms. However, we have designed the *Single-Template* baseline so that it mimics a traditional rule-based poetry generation algorithm, wherein a single POS template is followed (Colton et al., 2012). For each prompt word, a random template is selected and *Single-Template* generates text according to it. This experiment will highlight the advantages of adaptively choosing templates.

From Table 7, we see that the means of 5 metrics are significantly greater than 3, which means AMTC has a clear advantage over using a single template constraint. This makes sense, since AMTC allows *LimGen* to adaptively choose which template to follow. Though AMTC is easy to implement, we see substantial improvement over its predecessors. Lastly, the mean of relatedness is 2.95, but the p-value is not small enough to claim that *LimGen* is worse than *Single-Template*.

4.6 Experiment 3: *LimGen* vs. *Candidate-Rank*

Candidate-Rank beam search (Cohen and Beck, 2019) addressed the degradation of beam search performance when the beam size grows too large. It is simple to implement, and remains one of the best modified beam search algorithms.

From Table 8, the only statistically significant result is that *LimGen* outperforms *Candidate-Rank* with respect to sensibleness, which is due to the diversity fostering beam search MTBS. Since in our left-to-right limerick generation procedure, *LimGen* picks the next word that not only satisfies POS, meter, syllable and rhyming constraints but also flows naturally with the preceding lines, it is beneficial to maintain a diverse set of preceding partial lines to choose from. This ensures coherency and sensibleness in the output limericks. We can see the role of MTBS in fostering diversity more explicitly by counting distinct POS templates and by calculating the repetition (in terms of n-grams) within a fixed number of output limericks. For both *LimGen* and *Candidate-Rank*, a maximum of 360 sentences can be processed in parallel. We ran both methods 200 times (using 100 prompt words, each with one female and one male name). *LimGen* has an average of 27 different templates per $\sim$200 poem run, whereas *Candidate-Rank* only used 6 templates on average. For each run, to measure diversity, we randomly selected 50 limericks from the output set and calculated the ''mean popularity of each n-gram'' (e.g., 2-gram, 3-gram, 4-gram, 5-gram) in their last lines. Specifically, for each n-gram (n consecutive words) within those 50 last lines, we record its number of occurrences within those 50 lines. We then average all those recorded

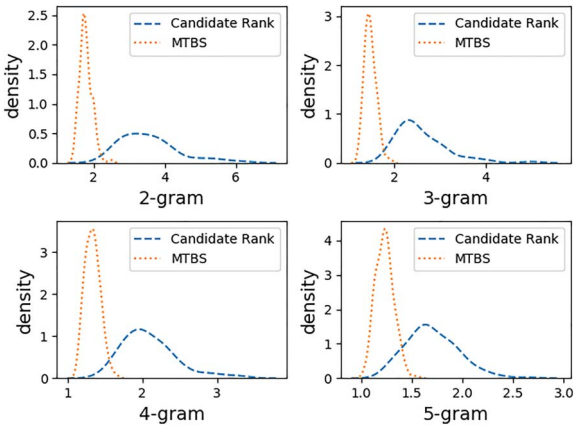

Figure 5: Distributions of ‘‘mean popularity of n-gram’’ within last lines for *LimGen* and *Candidate-Rank* output.

numbers and denote it as the ‘‘mean popularity of n-gram.’’ For instance, ‘‘mean popularity of 3-gram’’$= 2.0$ indicates that, on average, each 3-gram within those 50 lines repeats twice. A high value of the ‘‘mean popularity of n-gram’’ indicates heavy phrase repetition. As we can see from Figure 5, MTBS has a significantly lower ‘‘mean popularity of n-gram’’ than the Candidate-Rank beam search, which indicates more sentence diversity within MTBS’s output.

4.7 Experiment 4: *LimGen* vs. *Deep-speare*

Similar to GPT-2, *Deep-speare*’s language model was trained on 2000 limericks for 30 epochs until validation loss stopped decreasing using the optimal hyper-parameters provided by Lau et al. (2018). Since the pentameter model for stress and the rhyming model in the full *Deep-speare* are not guaranteed to adhere to limericks’ stress, syllable, and rhyming constraints, especially when the training data are scarce, we replaced these two models (pentameter and rhyming) with constraints to ensure the output from *Deep-speare* meets the requirements of limericks. Compared to GPT-2, *Deep-speare* is a much smaller model. In the original paper, it was trained on only 7000 quatrains of sonnets. After training on our limerick dataset, it was able to produce some form of limerick that warrants a comparative experiment.

We can clearly see from Table 9 that for the task of limerick generation, *LimGen* outperforms this adapted version of *Deep-speare* (which is considered a state-of-the-art neural network for English poetry generation) across all metrics. It remains to be seen whether *Deep-speare* will improve given more training data. However, it is unclear where more data would come from.

Metrics \ Statistics	mean	sd	p-value
emotion	**3.46**	1.18	2.96×10^{-7}
grammar	**4.02**	1.35	≈ 0
humor	**3.36**	1.24	6.74×10^{-5}
storytelling	**3.98**	1.11	≈ 0
sensibleness	**3.99**	1.18	≈ 0

Table 9: *LimGen* vs. *Deep-speare*.

4.8 Experiment 5: *LimGen* vs. Human Poets

In this experiment, 50 human limericks were chosen randomly from our database. Although not completely homogeneous in their poetic qualities, they were all well-thought-out and well-written, and represent genuine effort from their authors.

In Table 11, we added a column that records the percentage of limerick pairs with an average response > 3, that is, the percentage of *LimGen*’s limericks that are better than human’s on a specific metric according to crowd-workers. Clearly, human poets outperform *LimGen* on several metrics. It is not statistically conclusive which method is better with respect to grammar, presumably due to the template-guided approach that ensures grammatical correctness. Upon careful inspection, we noticed that for several metrics, there is actually a significant portion of *LimGen*’s output that were rated more highly than human-written limericks. For example, 43% of the machine-generated limericks had better emotional content than human poems. Another observation is that humor seems to be the hardest attribute for *LimGen* to emulate and master. Even though *LimGen* does output humorous limericks at times, they usually do not have the highest score according to our scoring function $H(\cdot)$; in other words, even though humorous poems were generated, our scoring mechanism could not recognize them as humorous.

In this same experiment, we asked crowd-workers a Turing test question for each limerick pair (one by a human and one by *LimGen*) (Figure 6): whether Limerick A or B is more likely to be written by a human. Recall that in our analysis we have transformed the data such that a score of 5 indicates the crowd-worker thinks that the poem was surely written by machine. The recorded score distribution is

There once was a brave soldier named Wade Who led a small army on his raid. He died on the campaign, His body burned again, But he kept his promises and stayed.
(**a**) Prompt word: war
There was a honest man named Dwight Who lost all his money in a fight. His friends were so upset, They were willing to bet, And they did not like feeling of spite.
(**b**) Prompt word: loss
There was a loud waitress named Jacque, Who poured all her coffee in a shake. But the moment she stirred, She was struck by a bird, Then she saw it fly towards the lake.
(**c**) Prompt word: shaken
There once was a nice man named Theodore Who killed all his family in a war. He came back from the dead, With a scar on his head, But he lost his memories and more.
(**d**) Prompt word: violent

Table 10: More Example limericks from *LimGen*.

Statistics / Metrics	mean	sd	p-value	> 3
emotion	2.84	1.41	0.04	43%
grammar	2.97	1.41	0.29	58%
humor	2.21	1.41	≈ 0	22%
storytelling	2.55	1.49	≈ 0	37%
sensibleness	2.58	1.47	≈ 0	35%

Table 11: *LimGen* vs. human poets.

$5 : 11\%, 4 : 14\%, 3 : 18\%, 2 : 29\%, 1 : 27\%$. Scores 4 and 5 are when *LimGen*'s limericks are mistaken as human-written when directly compared with actual human-written poems. Score 3 is when judges cannot differentiate between *LimGen*'s output and human poems. *Overall, the crowd-workers cannot differentiate LimGen's output from human-written poems 43% of the time.*

While so far we have compared *LimGen* with baselines and prior works on a relative scale because people are better at comparing items rather than assigning direct values to them, we now evaluate *LimGen*'s output on an absolute scale, which would paint a clearer picture of its strength and weakness on poetic metrics. We convened an expert panel of 20 Duke University students who are proficient in English, have received a liberal arts education and have completed two college-level courses designated to satisfy the literature requirement of the university. Since the intended audience of limericks is the general public, we believe that these panelists, with their considerable experience and expertise in the English language, are qualified to directly evaluate 60 limericks (30 from *LimGen* and 30 from humans) across the same metrics on an absolute scale from 1 to 5 (1 being the worst and 5 being the best). Each panelist completed at least one assignment, which consists of 6 poems randomly chosen from the set of 60 limericks. We ensured that each limerick was evaluated at least twice and the panelists did not see repeated limericks. None of these panelists knew anything about how the automated poems were generated. They were only notified that they would see a mixture of machine and human-written limericks.

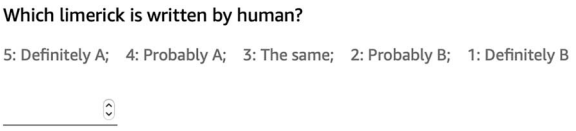

Figure 6: The Turing test question.

The scores in this survey are absolute values rather than relative values. We interpret an average over 3 on a metric as a decent level of performance. From Table 12, although expert judgment confirms that human poets outperform *LimGen*, it still shows that *LimGen* performs decently according to several metrics: *LimGen* has decent grammar and can tell a story well with its verses. It seems that grammar and storytelling are the easiest poetic attributes to master, since both human poets and *LimGen* have the highest scores on these metrics. Emotion and sensibleness are harder to learn. But what really differentiates human poets and *LimGen* is poets' ability to consistently make jokes.

Overall, we find our results encouraging, as they not only show that *LimGen* outperforms all prior baselines by a clear margin, but also shows that *LimGen* has the potential to approach human level performance in the future. More outputs from *LimGen* are in Table 10.

	Human	*LimGen*	p-value
emotion	3.79 ± 0.98	2.93 ± 0.99	0.006
grammar	4.22 ± 0.99	3.65 ± 0.96	0.068
humor	3.92 ± 1.01	2.21 ± 0.92	≈ 0
storytelling	4.44 ± 0.74	3.68 ± 0.85	0.009
sensibleness	3.88 ± 0.92	3.03 ± 1.05	0.006

Table 12: Expert judges: *LimGen* vs. humans.

There was a shy actor named Dario, Who played a big role on our show. He came back from the break, And we went to the lake, And he sat down and took his photo.
(**a**) Prompt word: Season
There was a artist named Cole, Who made a huge impact on my soul. He was a musician, He was on a mission, And that is the beauty of this role.
(**b**) Prompt word: Art
There once was a liar named Kai, Who fooled a grand jury on her lie. I had a suspicion, I was on a mission, I was ready to fight and to die.
(**c**) Prompt word: Cunning
There was a bright cleaner named Dot, Who put all her money in a pot. When she started to smoke, She was struck by a stroke, She has a severe case of a clot.
(**d**) Prompt word: Water
There was a funky chef named Dwight, Who cooked a great meal on our night. We got back from the bar, And we walked to the car, And we sat down and had our bite.
(**e**) Prompt word: Beer
There was a cruel judge named Lyle, Who killed a young girl on his trial. It was like a nightmare, I was scared by his stare, But I knew his intentions and smile.
(**f**) Prompt word: Death

Table 13: Additional limericks from *LimGen*.

5 Conclusion

LimGen is the first fully automated limerick generation system. Using human judgements, we have shown that our adaptive multi-templated constraints provide *LimGen* with a combination of quality and flexibility. We have shown the value of our diversity-fostering multi-templated beam search, as well as the benefits of our Storyline algorithm.

Acknowledgments

We would like to extend our sincere appreciation to all people involved in this research project, especially our colleagues Matias Benitez, Dinesh Palanisamy, and Peter Hasse for their support and feedback in the initial stage of our research. We would also like to thank Alstadt for funding. We have included a few more poems from *LimGen* in Table 13. Please refer to our online GitHub repository (Wang et al., 2020) for implementation details and more poems.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D17-1098

Doug Beeferman. 2016. Datamuse. https://www.datamuse.com/api/.

John Benhart, Tianlin Duan, Peter Hase, Liuyi Zhu, and Cynthia Rudin. 2018. Shall i compare thee to a machine-written sonnet? an approach to algorithmic sonnet generation. *arXiv preprint arXiv:1811.05067*.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*, O'Reilly Media Inc.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4925–4931. International Joint Conferences on Artificial Intelligence Organization. **DOI:** https://doi.org/10.24963/ijcai.2019/684

Sheila Cicchi. 2019. Internet limericks. https://www.brownielocks.com/Limericks.html.

Carnegie Mellon University (CMU). 2019. The CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1290–1299, Long Beach, California, USA. PMLR.

Simon Colton, Jacob Goodwin, and Tony Veale. 2012. Full-face poetry generation. In *Proceedings of ICCC-2012, the 3rd International Conference on Computational Creativity*, pages 230–238.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10696. **DOI:** https://doi.org/10.1109/CVPR.2019.01095

Pablo Gervás. 2000. 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*, pages 93–100.

Pablo Gervás. 2001. Generating poetry from a prose text: Creativity versus faithfulness. In *Proceedings of the AISB 2001 Symposium on Artificial Intelligence and Creativity in Arts and Science*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191. **DOI:** https://doi.org/10.18653/v1/D16-1126

Jim Haynes. 2010. *The Great Australian Book of Limericks*. Allen & Unwin.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546. Vancouver, Canada. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P17-1141

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://github.com/explosion/spaCy.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178. **DOI:** https://doi.org/10.18653/v1/P17-1016

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W17-3204.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958, Melbourne, Australia. Association for Computational Linguistics.

Edward Lear. 2010. *A Book of Nonsense*, Kessinger Publishing.

Gershon Legman. 1988. *The Limerick*, Random House.

Robert P. Levy. 2001. A computational model of poetic creativity with neural network as measure of adaptive fitness. In *Proceedings of the ICCBR-01 Workshop on Creative Systems*.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *CoRR*, abs/1601.00372.

Internet Limericks. 2019. Internet limericks. https://www.familyfriendpoems.com/.

Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images

by multi-adversarial training. In *ACM Multimedia 2018*. **DOI:** https://doi.org/10.1145/3240508.3240587

Ruli Manurung, Graeme Ritchie, and Henry Thompson. 2000. Towards a computational model of poetry generation. https://era.ed.ac.uk/handle/1842/3460.

Namepedia. 2019. Namepedia: The name database. http://www.namepedia.org/.

Eric O. Parrott. 1984. *The Penguin Book of Limericks*. Penguin Books.

PoemGenerator. 2019. Poem generator. https://www.poem-generator.org.uk/limerick/.

PoemOfQuotes. 2019. Poem of quotes. https://www.poemofquotes.com/tools/poetry-generator/limerick-generator.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://www.aclweb.org/anthology/N18-1119.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://github.com/openai/gpt-2.

Stanton Vaughn. 1904. *Limerick Lyrics*. J. Carey & Company.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing He Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 7371–7379.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1. **DOI:** https://doi.org/10.1109/TPAMI.2016.2587640, **PMID:** 28055847

Jianyou Wang, Xiaoxuan Zhang, Yuren Zhou, Chris Suh, and Cynthia Rudin. 2020. Online github repository for LimGen. https://github.com/wjyandre/LimGen.

Kai Wang, Xiaojun Quan, and Rui Wang. 2019. BiSET: Bi-directional selective encoding with template for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2153–2162, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1207.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060, Osaka, Japan. ACL.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1356

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. 2016. i, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2197–2203.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic Chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1430

Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou. 2020. Improving

neural machine translation with soft template prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5979–5989, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.531

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9450–9457. **DOI:** https://doi.org/10.1609/aaai.v34i05.6488

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018a. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153, Brussels, Belgium. Association for Computational Linguistics.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018b. Chinese poetry generation with a working memory model. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 4553–4559, AAAI Press.

Xingxing Zhang and Mirella Lapata. 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. **DOI:** https://doi.org/10.3115/v1/D14-1074