Fiction Sentence Expansion and Enhancement via Focused Objective and Novelty Curve Sampling

1st Yuri Safovich
Data Science Center
Ariel University, Ariel, Israel

2nd Amos Azaria
Data Science Center and Computer Science Dept.
Ariel University, Ariel, Israel

Abstract—We describe the task of sentence expansion and enhancement, in which a sentence provided by a human is expanded in some creative way. The expansion should be understandable, believably grammatical, and highly related to the original sentence. Sentence expansion and enhancement may serve as an authoring tool, or integrate in dynamic media, conversational agents, and advertising.

We implement a neural sentence expander, which is trained on sentence compressions generated from a corpus of modern fiction. We modify the objective loss function to support the task by focusing on new words, and decode at test time with controlled curve-like novelty sampling. We run the sentence expander on sentences provided by human subjects and have humans evaluate these expansions. The generation methods are shown to be comparable to, and as well liked as, subjects' original input sentences, and preferred over baselines.

I. INTRODUCTION

One of the most important skills acquired in elementary school as well as in high-school is the art of writing. While young authors are able to write short and simple sentences, they struggle with writing more complex ones [1]. This is because writing is quite different from speech, and it takes some effort for elementary or even high-school students to develop and write complex sentences. The task of sentence expansion and enhancement, which we describe in this paper, is to take a short sentence provided by a human or an agent, and expand it in some creative way, to a more sophisticated sentence. The new sentence should mostly preserve the content of the original but may not strictly contain the original text (as opposed to infilling). The enhancement may be in clarity, memorability or arbitrary subjective measures, and is distinguished in this task from other sentence expansions such as in textual entailment. Enhancement to a same-length or reduced-length sentence would, if applied, respectively parallel the tasks of paraphrasing and sentence compression, and while we design for expansion, some of our methods are applicable to these related objectives. The expansion is successful if the judge prefers it to the original in its context. That is, the task of sentence enhancement is subjective, and depends very much on human perspective

Sentence expansion and enhancement may serve as an authoring tool, for assisting authors to compose more complex sentences or converting a "summary" tell into show, or be integrated in dynamic media, such as games, creating more interesting interactions. Static, human-authored game dialogue

systems, for instance, can be likened to curated interaction with conversational agents, and may be an ideal target for a creative writing interface, such as the proposed, as these dialogue systems also do not require advance plot integration planning. Sentence enhancement can also be used by advertising, for creating a variety of options of conveying a specific message, and then picking one or more expanded sentences which are assumed to perform best (e.g., those not shown to this user previously). Finally, conversational agents and virtual assistants [2, 3, 4] and recommender system agents [5] may use sentence enhancement by generating only the essence of the sentence that needs to be communicated, and this sentence may in turn be expanded to be more human friendly and entertaining.

For example, given an input sentence such as "hello world", a generative model would produce for us "Oh, hello, a world of peace." (This is an actual model output). An ideal model would be better at the task than an average human, and this appears feasible, by our results.

Computational creativity has always been a popular idea; automatic storytelling and poetry have been attempted from early in computing. Narratology, the study of storytelling, divides it into story (plot) and discourse (style, chronology of presentation) [6]. Most research has been focused on the former [7]. For example, a set of elements and actions is given, with preconditions and postconditions, and then the construction of a story plot is a search problem. The use of deep neural networks permits generating the language in an integrative way with respect to plot elements. Without some direction, however, the outputs may lack innate meaning. An early example of fully abstractive generation via character-based language model recurrent neural networks (Char-RNN) [8]: "while he was giving attention to the second advantage of school building a 2-for-2 stool killed by the Cultures saddled with a half- suit defending the Bharatiya Fernalls office." This artifact is a human-readable metric, a language modeling result used for comparison in neural network research. Although more convincing results appear as the field develops, better grounding for a generative model may be attained by drawing on human participation for input. This enables more nuanced output than in pure generation.

In the literature, most narrative generation methods have been extractive, meaning chosen words or connections are present in some source schema, and logic-, graph- or template-

TABLE I
EXAMPLE SENTENCE EXPANSION AND ENHANCEMENT OUTPUTS FOR HUMAN INPUTS.

Input	the woman glared at the child.
Output	the old woman glared down at the younger child.
Input	the robot looked at jake and smirked, it seemed.
Output	the little gray crew looked out at jake and scowled, it seemed like a greatly bad-tempered one.

TABLE II
EXAMPLE KERNELS AND ORIGINAL SENTENCES.

Kernel	smoke belched from the pipe.
Original	blue smoke belched from the chromed exhaust pipe.
Kernel	are you back?
Original	are you back with the revolutionary lover?

based [9, 10, 11, 12, 13]. The topic of sentence enhancement would include slightly modifying words and concepts. In the context of deep learning, abstractive generation is easier, and may be interesting also for the potential relationship with general creativity in AI. With the recent advances in hardware and neural networks, interest grew in abstractive text generation models, due to capacity for generalization, long-range interactions, and to reduce manual knowledge modeling. A common problem with the essentially statistical systems is generation of safe sequences that are relevant for many inputs [14]. Fully abstractive generation may result in good-looking but meaningless or irrelevant text [8, 15]. Some have mixed neural components in extractive work, and relatively newer models, such as generative adversarial networks (GANs) and variational autoencoders, learn to generate text with diversity through additional parameters in generation or training. However, success relative to standard neural language models is debatable [16, 17] for tasks applicable to both. At inference, on the decoder side of encoder-decoder sequence to sequence (seq2seq) [18], beam search tends to produce more generic additions while random sampling is more unreliable, in particular for the task we propose.

In this paper, models are developed to obtain human input and expand it, transforming complete but possibly unembellished sentences to give them a general or specific style (see Table I). Given an input (*the robot looked at jake...*), the models add content or context to the input sentence. In terms of the narratological split to story and discourse, the inputs are story clauses. In the example, *robot* expands to *little gray crew* (member?) demonstrating abstractive generation.

We compose a corpus of story sentences paired with their compression ("kernel"). These sentence kernels are obtained by using sentence compression techniques on a corpus of modern (mid to late 20th century) fiction, which is scraped from online resources. See Table II for examples of compression kernels. Using this, we train models to transform sentence "kernels" to their source form, using the platform of seq2seq with attention [19] for the expansion task. The seq2seq loss function is first modified to emphasize learning new words, making extended training possible. Then simple alternative test-time sampling methods are applied to better control randomness in decoding. The changes improve output quality, in terms of the average preference of human judges.

We run the sentence expander system on crowdsourced human input. It is shown that the best method for sentence expansion results in sentences that are as well-liked by crowd-sourced human judges as the input.

To summarize, the main contributions of this paper are threefold. First, the problem of sentence expansion and enhancement is defined and its importance explained. Second, a method is presented that allows the generation of a large dataset for the sentence expansion problem, by using sentence compression techniques on a given corpus. Capturing the lost information from sentence compression is an approach which to the best of our knowledge is novel. Using compressed sentences does not require any labeled data, and thus allows the construction of data from any source. Third, a method is presented for automatic sentence expansion, which is based on several novel ideas, and it is shown that humans prefer sentences produced by the system to original input sentences significantly more often than with baseline systems.

II. RELATED WORK

A. Sentence Compression

Given an input sentence, sentence compression produces a shorter sentence preserving meaning. Text deletion-based (i.e. extractive) models, used extensively, and newer, abstractive models also, employ word and phrase substitution and re-ordering, learned from data. A commonly used corpus for abstractive compression or summarization training is Giga-word [20], which comprises ~10m documents (4b words) of newswire (with headlines) and auto-generated syntactic annotations. A CNN seq2seq with attention summarizer by Rush et al. [21], trained on Gigaword, first-sentence to head-line, in their evaluation outperforms an extractive ILP-based (integer linear programming) model by Clarke and Lapata [22] and other baselines, although this is likely due to the nature of newswire headlines. Toutanova et al. [23] compare various metrics, including human evaluators and using four compression systems, and report an opposite relationship on a multi-genre corpus wherein ILP is state of the art. ILP as used is a learning algorithm in one form but performance relies on linguistic constraints. Filippova and Altun [24] report a method for building a parallel corpus for extractive compression from news headlines and first sentences. Filippova et al. [25] use it to learn LSTM deletion sequences (left to right) with a 2m pair news corpus, reporting 30% (versus 20%) perfect match, showing that syntactic features are not required in these DNN models.

F'evry and Phang [26] show controllable-length neural compression without a parallel training corpus by denoising autoencoders, learning to reconstruct a sentence from a list of its words and, as noise, some from another. Desired length is a decoder input.

B. Sentence Generation

Generation of full text sentences from a mapping, as in translation, may be done with the familiar sequence to sequence (seq2seq) models, which are used in this work as well. Various decoding methods for diversity are found in the field of dialogue generation; however, not many are applicable to the task of expansion. A survey is in Ippolito et al. [27].

One example of non-neural controlled sentence generation is by Özbal et al. [28], a framework for slogan generation from keywords, using a dozen feature functions to select words filling slotted patterns. Some of the keywords are special, such as emotion and domain. This is in contrast to arbitrary specification, in language, as in our model.

On the other end, Fan et al. [29] collected a dataset of short ($\sim$700 word) stories written for a sentence premise. They emulate the premise with a GAN generating sentence-length prompts, and then generate stories using convolutional seq2seq, with training-time model fusion [30]. To support the output length they incorporate gated self-attention heads at different frequencies. This is intended to spread the prompt concepts over many sentences. We train their story expansion seq2seq model on our sentence dataset for a baseline.

Style transfer and conversational models generate sentences from sentences, the latter with context. Wang et al. [31] use an inverted objective and decode using an MLP sample selector. They focus on a topic by feeding a grid-based topic embedding to the decoder. Prabhumoye et al. [32] learn the sentence itself as a latent variable before adversarially generating against style classifiers. Su et al. [33] sample sentences in a MCMC process, with a discriminator for constraints such as sentiment. Of research that adapts training for the sampling, Ranzato et al. [34] optimize sequence decoding based on BLEU or ROUGE scores, using reinforcement learning. From a design perspective, Clark et al. [35] discuss aspects in creative writing with AI, based on slogan generation and story next-sentence suggestion experiments.

III. METHODS

A. Focused Objective

Using a simple seq2seq model to predict the original sentences from their kernel (the compressed sentence) results in a model that simply copies the input and does not expand it at all. This is because all the words that appear in a kernel appear in the original sentence (which is used as the expanded sentence) and merely learning to copy the input provides a relatively low loss. This is clearly not useful. Early stopping does not help as it produces an inadequately trained model. However, an incentive could be provided for the model to add additional words not seen in the input, but which appear in the expanded sentence. To do this, the negative log-likelihood loss function is modified to increase the importance of learning new words. The cross-entropy of every word in the target (expansion) that is not in the source is multiplied by a factor. After empirically testing values from 2 to 100, we chose 10 to balance effect and convergence. Let $\mathtt{I}$ denote the indicator function for a set, T the ground truth, S the source tokens, and $\lambda = 9$ (so that $1 + 9 = 10$ for the focusing factor). The modified cross-entropy is thus:

$$-\sum_t (1 + \lambda \mathtt{I}_{T-S}(w_t)) \log p(w_t | w_1, \ldots, w_{t-1}) \quad (1)$$

With this change, the model no longer degenerates to copying its input and can be trained for as long as desired. Preserving words of the input sentence becomes challenging for the model; however, this does not necessarily detract from the goal of improving the input. The output becomes longer and more diverse, even without random sampling.

B. Controlled Sampling

Random sampling is known to give diverse output. However, with softmax temperatures in the most favourable range for our dataset ($0.3 - 0.7$), in poor expansions we observe arbitrary digressions abruptly halted by the attention search, and on the other end failures of randomness degenerating to the argmax. As expressed in Table III, different sampling temperatures, beam search, and greedy decoding are empirically equivalent in performance.

Generally speaking, while composing sentences, authors may begin with an intention or idea (which we represent as a sentence kernel) and then decide where to take it or how to write it. Moreover, in a story, the principal plot twists, including the setting setup, tend to occur at significant distance from each other, in order to prevent the audience from becoming fatigued or wary. A conspicuous example of this is the periodicity of cliffhangers in some works. Other examples include intervening segments of comic relief, which may release tension without resolving plot points, and "filler" episodes, which pad length. Following this, we examine the concept of a novelty curve, in the simplified story of a single sentence.

We aim for a fixed degree of overall novelty in a sentence, so that it is not affected by sentence length in unstable fashion. In addition, controlling this enables use of models for the above concepts with a fixed count of loci. We calculate per-word novelty as the difference from the softmax maximum. That is, with p and τ as the probability and corrected sampling temperature at step t,

$$
\begin{aligned}
nov_w = \max_y \; &\mathtt{softmax}(\frac{\log p(y)}{\tau}) \\
&- \mathtt{softmax}(\frac{\log p(w)}{\tau})
\end{aligned} \quad (2)
$$

Then, novelty is rationed over expected length, using an accumulator and adjusting the sampling temperature. More complex methods than sampling may be substituted at this point, such as merging some of the top outputs with another layer. To control the rationing several adaptive curves were tested. Each is based on a model of the concept, e.g. novelty that is parabolic, exponential, cyclic, moving window, etc., and adjusts itself (in a step-based decoder) using previous sampling outputs towards a target overall sentence novelty.

The highest-performing model was parabolic. That is, with the middle of the sentence as the center at 0.5,

$$\tau = b^2(x - 0.5)^2 + c \tag{3}$$

This model adjusts the corrected temperature τ using Eq. 4, where on the left is an integral over τ for remaining temperature under the curve, with a as the time (current step) divided by expected length, and on the right t is the remaining novelty (target minus accumulated).

$$\int_a^1 \left(b^2(x - 0.5)^2 + c \right) \, dx = t \tag{4}$$

Solving for one of the parabola's parameters, b^2 or c, with the other set experimentally as a constant, gives two possible variants for this model, with similar performance.

$$b^2 = \frac{12(c - ac - t)}{4a^3 - 6a^2 + 3a - 1} \tag{5a}$$

$$c = \frac{-4a^3b^2 + 6a^2b^2 - 3ab^2 + b^2 - 12t}{12(a - 1)} \tag{5b}$$

To assist in aiming for fixed novelty, without retrying until it is within error, some tuning was done as follows. To prevent novelty "spikes" at parabola ends unintentionally exhausting the quota too early, top-40 sampling is used, which also reduces irrelevant generation. This can be compensated for by using slightly more powerful parabolas than calculated. For the listed parabolic curves the free hyperparameter is adjusted at design time; other curves can be scaled as needed. UNKs are not generated. To reduce repetitiveness in a simple way, repeated words in a 5-token history are penalized. The settings used are $b^2 = 0.5$ and $c = 3$ for the respective parabolas, repeat penalty is 15 for content words and 10 for stopwords, and the projected expansion factor is 1.65, which is approximately average for both test and corpus expansions. Computationally, these additive steps and the locally contained formulas for temperature resolve in time comparable to random sampling.

There may be other parabolic models, or models following other rules, that would be performant. Some less successful models in tests include an exponential on remaining novelty (up to expected length), which spikes τ somewhere inside the sentence, determined by a coefficient. Another is a windowed accumulator, of 3 or 5 tokens, balancing novelty, following either the target novelty or a parabola. The latter illustrates several concepts in Fig. 1. As shown, τ, the calculated temperature starts high and words are chosen such that each diverges more from the original sentence: shortly, off, my. After 4 tokens τ drops as it enters the center of the parabola at half the expected length, producing words that connect the digression with the original sentence: left, comma, the. Near the shallow center *tree* would be a likely token, and would separate the first novelty spike from others. However, in this case τ quickly rises again producing the words huge, bare, beige, Christmas. At this point the curve is truncated to a limiting value $\epsilon = 0.1$. The value in orange shows the accumulated novelty over the

past 3 tokens. The accumulator's effect is weak in this case and model, but can be seen on the parabolic curve at steps 4 and 7, where respectively it decreases or increases τ to compensate for too much or too little novelty.
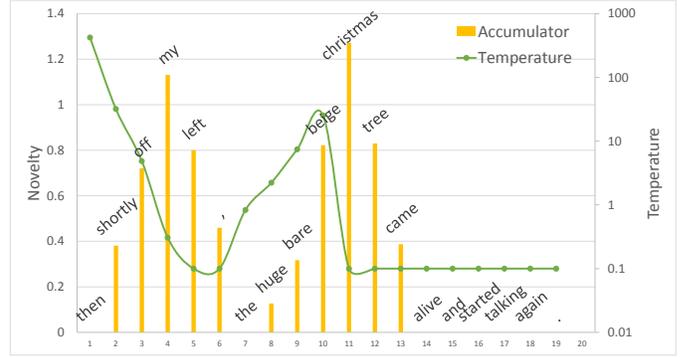

Fig. 1. Decoding the human input "the tree came alive and started talking" to get "then shortly off my left , the huge bare beige christmas tree came alive and started talking again". Zoom in for outputs. This instance uses a truncated constant parabola model modified by a size-3 window accumulator (value in orange). Corrected temperature τ is in green.

IV. EXPERIMENTS

A. Corpus

Large public corpora of English fiction (e.g. Google Books, Internet Archive) have well-known quality issues with formatting, optical character recognition, and content categorization. One exception is Project Gutenberg, which is proofread. Project Gutenberg's 19th and early 20th-century public domain fiction and nonfiction has dated English, which transfers noticeably in our model. Another corpus used by some authors is BookCorpus, a dump of free user-created fiction from one source. Like several other corpora in use, this has biases, with most of the works being from the romance genre.

For this paper, a corpus was assembled by scraping the Internet for posted content; specifically, identifying and collecting proofread 20th century English fiction by published authors. We believe that this type of cross section better represents the learned Western experience, and can be valuable for sensitive generation tasks. It is hoped that other investigators are likewise inspired to recreate mutable but representative datasets, which are scarce in independent research. Our collection has approximately 600m words in 41m sentences, 45% of which is speculative fiction.

B. Optimization

In attempts at improving the quality of training, including to reduce the tendency for diverging phrase interpretations and expansions, the corpus was experimentally split into groups by topics. One method used was K-means clustering with a bag-of-words approach. Sentences were word-stemmed and vectorized by either TF-IDF (in this scenario, how specific a word is to its sentence) or hashing. Latent semantic analysis [36] was optionally employed at several dimension parameters (50,100,200,300). The silhouette coefficient of cluster

cohesion (defined as the average of scaled point distances to nearest different cluster) was highest (0.621) in the case with 10 clusters (for "genres"), with LSA to 200 components, 10k features, and counting words showing in 0.001% to 1% of sentences. In most cases the clusters were moderately self-similar in appearance and often could potentially be classified as, for example, "military", "bar/pub", "anatomy", or "Star Wars". Such clusters can be used as scenting sets for style priming. Unfortunately, in all cases one cluster was much larger than all others. As the clusters could not be balanced, this approach could not be used to split the corpus. Domain adaptation via other methods, such as in Axelrod et al. [37], is an avenue for future work.

In a different approach, genre qualifications embedded in the corpus were considered. The romance genre, 7% of corpus, may be considered fairly homogeneous and a model was trained on this subset, but was not found competitive. Outputs were significantly and perhaps unsurprisingly colored by a focus on relations between objects (whether spatio-temporal or social); this however meant there were fewer novel "idea" objects introduced, conflicting with the rationale for abstractive generation, as well as reducing interest for other genre inputs. Since this would limit the possible forms of output, the genre subset was not used further.

C. Setup

1) Compressor: To generate sentence kernels, the ILP-based system by Clarke and Lapata [22] is used. A standard KN-smoothing LM [38] with $1e{-}7$ pruning was used for the compression. A supervised model [21] was also tested using released data. Given the highly restricted nature of news prose which it is trained on, however, in its published configuration this system does not summarize fiction-style sentences convincingly. Inspection revealed a tendency to force a geopolitical framing, and a misunderstanding of common sentence structures in fiction.

2) Data: The corpus was cleaned of outliers and languages beside English using stopwords and inspection. Text was extracted and preprocessed to segmented sentence form by custom tokenization and segmentation, followed by CoreNLP [39]. A large number of exceptional cases in punctuation or style across time, authors, and proofreaders requires that the process is imperfect and some output with excessive-hyphenation was seen, among other issues. In the implementation, neutral punctuation, especially quotation marks, often compressed incorrectly. 10% of sentences are in quotes; 3% of test set sentences have quotes outside words. Quotes were consequently removed; however, a model trained with them does generate dialogue and narration together. Target compression was set to 40%; average was 31%. A subset was selected where at least 30% reduction occurred, in 17 million sentences. The use of this set corresponds to the technique of separating short items from a neural model, and has similar observed advantage. This subset is the base for training. 3000 sentences were held for development. Sets were shuffled and lowercased, digits replaced by #.

3) Training: Models with 4 layers LSTM 1024 encoder and decoder were trained for 1 million steps with 0.2 dropout. Vocabulary size is 50k using SentencePiece [40]. We note that with many neologisms and domain terms in science fiction, a common genre in this corpus, many "literary" words may not appear in a 100k regular vocabulary. Names were not removed, in order that they may be synthesized directly (and thematically). A subword BPE vocabulary [41] tended to produce many nonsense words in experiments; SentencePiece produced less. Sentences over 50 words (1.5%) had words in excess truncated.

4) Test: A test set of 100 sentences was crowdsourced from 20 workers on Mechanical Turk. Workers were asked to author a "short sentence that might have appeared in some imaginary story", with no example given. Statistically, lengths are balanced (with a mean of 12 words, standard deviation 5.2). The mean was affected by the size or width of workers' input text area in multiple rounds of collection. Sentences with profanity or political entities were filtered.

During model evaluation, each input sentence and its expansion (obtained from the model) were compared by 3 unique, high-quality US workers, which were to choose the sentence that they "think is better or more interesting". The instructions to the workers were intentionally as short as possible, so that their subjective perspective is captured. The workers were not provided any example answers, to minimize researcher bias and let workers weigh brevity against novelty. That is, workers had to select either the input sentence or its expansion, separately for each model. Workers were not asked to rate entailment, but automatic entailment was computed, as described below. Comparisons were shuffled and presented in groups of 5 per worker (for reliability analysis), with random selection order. The workers were *not* instructed that one of the sentences is the initial and one is the expansion (as doing so would clearly bias their responses), even though this would likely be inevitable if paragraph context were given.

Expansions that failed to terminate within the length limit (50 tokens) or have clearly unnatural repetitiveness, defined as a sequence of length ≥ 10 repeated within a distance of 15, were removed (this happened in about 2% of the sentences for our method, but more sentences were removed for the baseline methods). The number of removals was reduced when penalties for repetitiveness were set.

D. Results

Table III presents the main results in terms of human preferences of the expanded sentence over the input sentence. In all models, excepting the kernel vs. original test, the input sentences were the 100 sentences written by the MTurk workers. It is also noted that only in one case a metric reached statistical significance ($r > 0.2$) for a sampling method, Frechet distance (described below) for random sampling. The term *Parabola* refers to methods arising from Eq. 4, with c or b^2 referring to the variable solved for in Eq. 5. Other custom curves did not perform as well and are not shown.

Baselines include:

TABLE III
SAMPLING METHOD EVALUATIONS.

Sampling	Preference	Significant metrics (if any)
Parabola (c)	**0.5**	
Parabola (b^2)	0.483	
Greedy	0.422	
Random 0.7	0.417	Frechet $r = 0.26$
Random 0.3	0.417	
Beam search	0.413	
Fan et al. s2s	0.3	
Trigram freq.	0.1	
Kernel vs. original	0.717	

1) Using the modified objective: random sampling with specified temperature, beam search (width 10), and greedy search;

2) Kernels held out from training, with original "expansions" (Table VI);

3) Inserting a word by sampling LM trigram frequency, up to average rate of expansion by other methods;

4) The seq2seq fusion model of Fan et al. [29] trained on our dataset for 500k steps, with outputs pruned of repeats in the same way. This is using the default top 10 sampling with temperature 0.8, comparable to other baselines; nonetheless, output length and diversity were relatively significantly random.

As depicted in the table, the human subjects preferred the original sentences (obtained from the original stories) to the compressed sentences (the kernels) 71.7% of the time. This is in fact our human-level upper bound, as the expanded sentences were actual story lines. Our parabola method outperformed the baselines, and appears to have reached human level equivalence in this scheme, with 50% of human subjects preferring the expansions to the original human input.

Example expansions for human input sentences, with human preference data, are given in Table IV. These examples are chosen to compare across methods, and illustrate user preferences, which are difficult to predict. Unfiltered examples are given in Table V. Some compression kernels and original sentences of writers (from the corpus) are given in Table VI.

1) Metrics: Automatic metrics that were tested have low Pearson's r and Spearman's ρ with evaluator preferences ($|r| \leq 0.1$). This varied across sampling methods but generally not to the point of significance.

The metrics computed were:

(i) Discrete Frechet and cosine distances in InferSent unsupervised sentence embeddings [42];

(ii) ratio of unique added unigrams and bigrams to length (Dist-1 and Dist-2 [14]);

(iii) ROUGE-1, ROUGE-2, BLEU-2, BLEU-4 [43, 44];

(iv) expansion ratio, added words, and input and output lengths.

No significant differences in variance or mean were observed in subranges upon plotting of embedding distances and other metrics. Using InferSent, reference MLPs [45] were trained on SICK dataset entailment and similarity [46] and

SNLI dataset entailment [47], and again r was negligible. Training the MLPs on preference data for prediction obtained $r = 0.02$ on a 10% test set. Additionally, a relation was manually evaluated on the (100-sentence) beam search results subset, as its relatively generic output may extend to the preservation of meaning (Table VII). There, for the accurate preservation of entities or concepts (58% of expansions) $r = 0.21$, and for strongly contradicting or changing the meaning (20%) $r = -0.18$.

V. DISCUSSION

Many definitions exist for creativity, and this can be source of friction for evaluation methodologies. In one definition, creativity is creating something novel and relevant (for someone or everyone). However, both *novel* and *relevant* may be task-dependent and subjective (for example, are all humans polled? From what age?). Our definition may include utility, while in another field creativity is subsumed by it. Hence, we seek to sidestep the issue of definition by minimizing our own involvement in query interpretation, passing the power to human observers. Then, the most that can be said is that the phrasing "better or more interesting" invokes different perception than a simple "prefer", which might focus more on grammaticality.

It has been argued that a blind evaluation incentivizes generators that are naive of context [48]. While we do not ask testers about relevance, in an experiment (not shown), automatic entailment was used to filter result data for those choices where entailment exists, from expansions to original sentences. In this case, the same relative ranking of most generation methods is arrived at, suggesting that their selection does not significantly influence context learning.

Since the task is very subjective in nature, it should not be surprising that raters had some disagreement on 63% of items in evaluations (on average), with chance-adjusted statistics suggesting low reliability in sets. The objective for testing was an untouched, unpartitioned measure of quality, and more agreement in low-performing model evaluations would possibly indicate a more uniform set of evaluators. Additionally, distinct NLP tasks exist for other possible elements that may be optimized for, such as sentence compression for brevity.

For sentence expansion, a more relevant or diverse output is not necessarily better. Presumably, each evaluator has different expectations from an expansion, and learning these is important for an authoring tool. The automatic metric results illustrate the necessity of early human evaluation. Preference data given in Table IV shows that preference can be counterintuitive, perhaps due to the diverse population of MTurk workers. In one experiment, two baseline methods, Random 0.3 and trigram frequencies, were compared, and the former was preferred 68.3% of the time, which may be less than expected given the latter's performance in Table IV.

Expansion outputs sometimes contradict, and the frequency of this is not explained purely by decoding method and the compression removing negators. Given the test collection methodology, input phrases might be inclined towards cliché,

TABLE IV

EXAMPLE EXPANSIONS FOR HUMAN INPUT, WITH PREFERENCE VS. INPUT (HUMAN EVALUATION).

Input	they were creeping around the corner when they heard a horrible scream.
Beam search	and then they were rushing around the corner, when they'd first heard a faint scream, and then turned to look at each other's eyes. (33%)
Random 0.7	and now, in all these other respects, they were both rushing back around the corner, when they'd first heard a strange scream of distress, and then very quietly. (0%)
Input	there was a princess that lived in a castle.
Beam search	but there was also a princess that still lived in a small castle. (33%)
Random 0.7	but there was also a romance-a queen that lived in a larger, spacious, rambling castle. (66%)
Exponential	now, there was a new, high-ranking and female-american soul that lived in a castle. (66%)
Input	she often wondered about it but she did not ask him.
Parabola c	she'd always wondered about it, but in that case she was not even to ask him. (66%)
Fan et al. s2s	but she'd often wondered about it, but she did not mean him in the first place, but in that kind of way. (33%)
Input	the kind found his perfect princess.
Fan et al. s2s	the kind of family found his perfect princess, the only one. (33%)
Beam search	the kind of man who 'd found his own, was a good catholic princess. (66%)

TABLE V

UNFILTERED EXAMPLE EXPANSIONS FOR HUMAN INPUT FROM THE PARABOLA c METHOD.

smoke was going into the room.
but this white smoke was going down into the empty room anyway.
why he was called will, she did not know.
why... he – therefore was called will, and then she did not even know how many times that would be in the first place.
the wizard gave a wry smile at his apprentice.
the dark old lord gave a deep, evil, almost imperceptible female.
call me jack.
please... call me, jack.
this is not her family's crest.
this is certainly not her family's crest, but the other two of my men aren't a good one.
she was the luckiest mom in all of the land.
at this, however she was the most beautiful, ungrateful, and stupid-and yet happy little kid in all of the land.
you are laughing at me, right.
but if this is true, then you are laughing at me from this point of view of yours, right.
i want to be a pilot.
i do want more than a few minutes, and to be here, a pilot.

TABLE VI

EXAMPLE KERNELS AND ORIGINAL SENTENCES, WITH PREFERENCE VS. INPUT (HUMAN EVALUATION).

Kernel	he put a hand on gabriel's shoulder and guided him.
Original	he put a hand on gabriel's shoulder and guided him from the kitchen and into the shadows of the yard. (100%)
Kernel	he has feeling for others outside circle of friends and attaches value to life.
Original	he has little feeling for others outside a very small circle of friends , and attaches little real value to human life. (33%)

TABLE VII

BEAM SEARCH (WIDTH 10) WITH MANUALLY EVALUATED ENTAILMENT AND INFERSENT DISTANCES.

r	Preserving (58%)	Contradicting (20%)	Frechet	Cosine	Dist-1
Preference	0.21	-0.18	-0.1	0.18	-0.12
Preserving	-	-	-0.24	0.48	-0.35

TABLE VIII
EXAMPLE OUTPUTS, GENERAL AND PRIMED STYLE.

Input	he woke up .
Style: *general*	he woke up in the brush.
Style: *Douglas Adams*	he woke up, carefully.

while in the corpus cliché is much likelier to appear in a subverted form. Conversely, conceptually dense sentences such as adages or the already published writing of a veteran author are unlikely to gain from extension (in general style). Splitting off coherence from meaning has not been part of our goal, but grammar and coherence remain significant factors positively correlated with user evaluations in our experiments, whether in input or expansion. This is also reflected in user feedback.

In Table VII there is correlation of manual entailment with InferSent distances ($r = 0.48$ for cosine) on beam search. If an entailment metric is considered reliable, it is easy to resample the output until entailment occurs; this does not seem to affect human preference, however.

Finally, the system allows narrowing the possible style as desired to some degree, using network bias, priming or scenting a pre-trained model with one author's books prior to decoding. A tiny sample is in Table VIII, although generation with specific author styles is not evaluated by humans here.

VI. Conclusion and Future Work

We have defined the task of sentence expansion and enhancement; the task includes a sentence input from humans and requires the production of a more literary, abstractively expanded sentence as output. The task is relevant in writing aids, where it may save time and potentially improve quality. Since content is added as well as style, it is relevant for virtual agents and assistants, games, text ads (adding variety), and other media benefiting from adaptable content. A parallel corpus of fiction sentences and their compressions is created, and models are trained on the reverse to perform expansion. A modification to the objective function encourages learning output features and makes training at nontrivial length possible. Simple curve-based sampling methods distribute output novelty in a controlled way. The model outputs, while not independently superior to human inputs, are shown to achieve parity in our evaluation scheme, reaching 50% of total expansions being preferred, out of an upper bound of 72% for, essentially, expert human expansions. The result surpasses baselines by 20% (compared to a sentence adaptation of the model by Fan et al. [29]). 12% of increase is obtained by the modified seq2seq, and a further 8% from controlled sampling, in the best performing method. Lastly, it is observed that common metrics of text generation do not predict user preferences for this task.

Our expansions did not compete directly with human generated expansions, and based on experiments the added uncertainty may contribute to a less stable metric, but this second context for comparison across methods, with user post-processing (editing of generated expansions), will be helpful

in proving that use of an expander indeed improves quality by saving time. In estimates from our experiments, untrained MTurk writers average one minute to expand a sentence in the same way and to the same length as our system. A complete evaluation suite would include building an assistive user interface, allowing users to choose between different sampling methods and to edit resulting sentences. The interface may learn personal preferences, used as feedback to improve future suggestions. Further developments of process may be expected for sentence expansion with paragraph context, and for paragraph expansion and enhancement. Evaluating alternative platforms or improvements to tools in this work is a logical step to surpassing human equivalence in this task setup.

References

[1] T. Hilfman, "Can second grade children write more complex sentences?" *Elementary English*, vol. 47, no. 2, pp. 209–214, 1970.

[2] A. Azaria, J. Krishnamurthy, and T. M. Mitchell, "Instructable intelligent personal agent." in *AAAI*, vol. 4, 2016.

[3] M. Chkroun and A. Azaria, "Lia: A virtual assistant that can be taught new commands by speech," *International Journal of Human–Computer Interaction*, vol. 35, no. 17, pp. 1596–1607, 2019.

[4] A. Azaria, S. Srivastava, J. Krishnamurthy, I. Labutov, and T. M. Mitchell, "An agent for learning new natural language commands," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 1, p. 6, 2020.

[5] A. Azaria and J. Hong, "Recommender systems with personality," in *Proceedings of the 10th ACM conference on Recommender systems*, 2016, pp. 207–210.

[6] V. Propp, *Morphology of the folktale.* Leningrad: Academia, 1928.

[7] B. Li, M. Thakkar, Y. Wang, and M. O. Riedl, "Storytelling with adjustable narrator styles and sentiments," in *ICIDS*, 2014.

[8] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *ICML*, 2011.

[9] B. Li, S. Lee-Urban, G. Johnston, and M. O. Riedl, "Story generation with crowdsourced plot graphs," in *AAAI*, 2013.

[10] R. M. Young and J. D. Moore, "Dpocl: A principled approach to discourse planning," *CoRR*, vol. abs/cmp-lg/9406020, 1994.

[11] M. O. Riedl and R. M. Young, "Narrative planning: Balancing plot and character," *J. Artif. Intell. Res. (JAIR)*, vol. 39, pp. 217–268, 2010.

[12] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *ACL*, 2008.

[13] ——, "Template-based information extraction without the templates," in *ACL*, 2011.

[14] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," in *HLT-NAACL*, 2016.

[15] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating sentences from a continuous space," in *CoNLL*, 2016.

[16] S. Semeniuta, A. Severyn, and S. Gelly, "On accurate evaluation of gans for language generation," *CoRR*, vol. abs/1806.04936, 2018.

[17] O. Cífka, A. Severyn, E. Alfonseca, and K. Filippova, "Eval all, trust a few, do wrong to none: Comparing sentence generation models," *CoRR*, vol. abs/1804.07972, 2018.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[20] C. Napoles, M. Gormley, and B. V. Durme, "Annotated english gigaword," in *Linguistic Data Consortium*, 2012.

[21] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *EMNLP*, 2015.

[22] J. Clarke and M. Lapata, "Global inference for sentence compression : an integer linear programming approach," *J. Artif. Intell. Res. (JAIR)*, vol. 31, pp. 399–429, 2008.

[23] K. Toutanova, C. Brockett, K. M. Tran, and S. Amershi, "A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs," in *EMNLP*, 2016.

[24] K. Filippova and Y. Altun, "Overcoming the lack of parallel data in sentence compression," in *EMNLP*, 2013.

[25] K. Filippova, E. Alfonseca, C. A. Colmenares, L. Kaiser, and O. Vinyals, "Sentence compression by deletion with lstms," in *EMNLP*, 2015.

[26] T. F'evry and J. Phang, "Unsupervised sentence compression using denoising auto-encoders," 2018.

[27] D. Ippolito, R. Kriz, M. Kustikova, J. Sedoc, and C. Callison-Burch, "Comparison of diverse decoding methods from conditional language models," in *ACL*, 2019.

[28] G. Özbal, D. Pighin, and C. Strapparava, "BRAINSUP: Brainstorming support for creative sentence generation," in *ACL*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1446–1455.

[29] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *ACL*, 2018.

[30] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," in *Interspeech*, 2018.

[31] D. Wang, N. Jojic, C. Brockett, and E. Nyberg, "Steering output style and topic in neural response generation," in *EMNLP*, 2017.

[32] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, "Style transfer through back-translation," in *ACL*, 2018.

[33] J. Su, J. Xu, X. Qiu, and X. Huang, "Incorporating discriminator in sentence generation: a gibbs sampling method," *CoRR*, vol. abs/1802.08970, 2018.

[34] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *CoRR*, vol. abs/1511.06732, 2016.

[35] E. Clark, A. S. Ross, C. Tan, Y. Ji, and N. A. Smith, "Creative writing with a machine in the loop: Case studies on slogans and stories," in *IUI*, 2018.

[36] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman, "Indexing by latent semantic analysis," 1990.

[37] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *EMNLP 2011*, 2011.

[38] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *ICASSP*, 1995.

[39] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[40] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[41] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *CoRR*, vol. abs/1508.07909, 2016.

[42] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *EMNLP*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 670–680.

[43] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," 2004.

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[45] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," *arXiv preprint arXiv:1803.05449*, 2018.

[46] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A sick cure for the evaluation of compositional distributional semantic models," in *LREC*, 2014.

[47] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *EMNLP*, 2015.

[48] A. Pease and S. Colton, "On impact and evaluation in computational creativity: a discussion of the turing test and an alternative proposal," 2011.