Creating Six-word Stories via Inferred Linguistic and Semantic Formats

Brad Spendlove and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
brad.spendlove@byu.edu, ventura@cs.byu.edu

Abstract

High-quality human-created artifacts are distinguished by cohesion and semantic richness that can be difficult for computational systems to emulate effectively. Certain classes of artifacts feature relationships between their constituent elements (e.g. words in a story) that naturally form a hierarchical structure that underpins the artifact's meaning. We formalize a tractable method for programmatically extracting such artifacts' structures framed in hierarchical Bayesian program learning (HBPL) and present HIEROS, a creative system that uses linguistic and semantic structures—collectively called *formats*—extracted from human-written six-word stories to guide the creation of novel stories. We describe how HIEROS infers formats from exemplar stories, how those formats inform its selection of words when writing novel stories, and how the system evaluates its stories to search for high-quality output. We present and evaluate stories HIEROS has written and discuss how our format model enhances the system's creativity.

Introduction

Certain classes of creative artifacts are built atop a latent hierarchical structure in which the artifact's constituent parts influence one another to various degrees. Examples include scenes in a movie, chapters in a novel, or notes in a song. We will refer to this class of artifacts as *hierarchically structured artifacts* that consist of a sequence of elements, with the relationships between those elements forming the edges of an acyclic hierarchy graph.

In addition to informing the artifact's form and meaning, this underlying hierarchical structure can yield insight into the processes by which the artifact was created. The ordering of the elements in a given hierarchy mirrors, to some extent, the creative process of a human creator. Altering the relative importance of these concepts focuses the creative process in different ways. For example, a songwriter may start with a melody then build harmonies or start with a certain chord progression and write a fitting melody.

We consider English-language stories as an example class of hierarchically structured artifacts whose constituent elements are words. In particular, we will examine six-word stories, a genre of microfiction that attempts to tell a meaningful or interesting story in just six words. The brief nature of six-word stories presents an interesting creative challenge and facilitates modeling and discussion. Although stories are the focus of this paper, any artifact that can be described in terms of a hierarchy of relationships between its elements can be analyzed using the approach presented here.

The words in a story artifact have hierarchical relationships with one another. The words are related to or influence each other both linguistically, such as in tense agreement between nouns and verbs, and semantically, such as in whether a certain pairing of words makes logical sense or conveys an intended action or mood. Not all of these relationships carry the same importance to the resulting story, however. The set of relationships between words in an artifact can be arranged into a hierarchy of dependencies ordered by importance. The exact meaning of "importance" is left purposefully vague here. Many valid hierarchical constructions exist for any artifact which may assign different importance to inter-word relationships or feature different relationships altogether.

For example, we can consider an underlying hierarchical structure to the story "the dog runs" in which the second word in the story is the most important and informs the first and third words in different ways. The relationship between the second word and the first is likely not as strong or important as the relationship between the second and the third. Furthermore, there are many ways to interpret the relationship between the second and third words, such as that they are merely a noun and a verb, that they are a subject and an action that that subject could take, or that they are a subject and action that specifically convey energy or urgency.

Note that the ordering of the hierarchy may not correlate to the sequential order in which the words are arranged in the artifact. Words form the artifact, and the relationships between them form the hierarchy. Other words could fit those same relationships, resulting in a different story but adhering to the same hierarchy. The story "a whale breaches" could be said to have the same underlying hierarchical structure as "the dog runs", such as the structure of a story featuring an animal subject, an article preceding that subject, and a verb that reflects the subject's energetic movement.

As language and its meaning are subjective, there may exist many possible hierarchical structures underlying any given artifact. The more a given hierarchy captures the interesting properties of the artifact and its meaning, the more

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

useful that hierarchy is. Another example story "the dog cowers" could be considered to have the same relationship between between 'dog' and 'cowers' as the previous example did between 'dog' and 'run', but there may be an important semantic difference between 'runs' and 'cowers' that a more useful hierarchy could reflect.

If a computational system can model an artifact's hierarchical structure, it can, by extension, model the whole suite of artifacts that fit in that same structure. A method for programmatically extracting useful hierarchies from creative artifacts would allow a computational system to model the relationships between an artifact's elements, thereby identifying the key components from which the artifact's meaning and quality are derived.

In this paper, we present HIEROS a computationally creative system for writing six-word stories based on a hierarchical model of story structure. HIEROS infers the hierarchical structures of human-written exemplar stories by constructing *formats* that are used to guide its creation of novel stories. We present the underlying model of story hierarchy that HIEROS implements, describe the system in detail, and analyze its results.

Related Work

While our approach to inferring formats for story generation is novel, previous creative writing systems also involve elements of constrained generation and learning from human-written artifacts. The models underlying many of these systems can be usefully viewed as hierarchical (or at least relational), allowing for more direct comparison to our system. In this section, we refer to several such systems and contrast them to the novel techniques used in HIEROS.

The storytelling systems MEXICA (Pérez y Pérez and Sharples 2001), STellA (León and Gervás 2014), and Fabulist (Riedl and Young 2006) all seek to guide the creation of story artifacts by imposing useful restrictions on the space of possible story events. These restrictions represent a hierarchy of probability distributions conditioned on previous events. HIEROS explicitly models such hierarchies and infers them from human-written exemplars instead of drawing from hard-coded concepts.

Both Colton et al. (2012) and Toivanen et al. (2012) present creative poetry-writing systems that model an exemplar's syntactic structure as a template for creating novel artifacts. The former system constrains the populating of the template with rules regarding the relationships of possible words with one another, while the latter models semantic relationships in a separate corpus to guide creation. These systems use fixed sets of constraints that are applied to all exemplar-generated poems, while HIEROS infers semantic constraint information from exemplars which are directly incorporated into its formats.

MICROS (Spendlove, Zabriskie, and Ventura 2018) presents a prototypical exploration of latent hierarchical structure in its automatic creation of six-word stories. The MICROS system creates story artifacts that are built upon an underlying structure, demonstrating the power of this model in guiding the creative process. MICROS' single, static story structure allows it to draw on powerful semantic knowledge bases but also heavily restricts its output variety. HIEROS shares MICROS' underlying model but improves upon MICROS by removing its fixed format restriction. Removing this restriction requires a more generalized method of generating and scoring stories and a method for automatically inferring templates from exemplars. The result we present here is a system that can generate stories according to a much broader set of formats than MICROS can.

Li, Wu, and Lan found that augmenting existing hidden variable models with syntactic and semantic structures improved their performance at machine comprehension tasks, demonstrating how modeling such structures is useful outside the domain of creativity (2018). We argue that our model is useful for any domain which concerns hierarchically structured artifacts, not just story artifacts. Any domain that permits viewing its artifacts as composed of elements and the relationships between those elements could be framed in this model for use in tasks not limited to generation.

Frame semantics (Fillmore and Baker 2001) is a linguistic theory that views a concept as a "frame" consisting of an arrangement of elements that comprise that concept, with each element representing a set of words that fulfill a specified role in that concept. Our model's formats bear some similarity to semantic frames in that both structures place restrictions on which words are acceptable to communicate a story or a concept, respectively. One implementation of frame semantics, FrameNet (Fillmore, Wooters, and Baker 2001), consists of a database of hand-annotated sentences and their resulting semantic frames. Our system seeks instead to infer formats automatically from human-written exemplars, allowing it to draw on a broader set of formats for generation.

Modeling Latent Hierarchical Structure Under HBPL

In this section, we develop a formal model for the general hierarchical structure of six-word stories. Similar models may be constructed for all types of hierarchically structured artifacts, with their constituent elements replacing words as the atomic units whose relationships the hierarchy captures.

Although the meaningful relationships between words in a story artifact may involve complex, recursive connections, in our model we simplify this hierarchy to a directed acyclic graph with at most one edge between any two elements.

If a directed edge exists from element p to element c, we refer to p as c's parent and c as p's child. We represent an artifact's hierarchy graph as a joint probability distribution over all possible words factorized into a series of subdistributions in which each word is conditioned only on its parent word in the hierarchy. Each subdistribution thus captures a relationship in the hierarchy by assigning high probabilities to words that follow from the parent word according to that relationship.

Thus, for the story "the dog runs" in which the hierarchy graph includes an edge (i.e. relationship) from the second word to the third word, 'runs' can be thought of as being drawn from a distribution of possible words conditioned on

the parent word 'dog'. If the relationship represented by this distribution is that the third word is an action that the second word could take, then other high-probability words in the distribution could be 'barks' or 'sits'. However, if the relationship is instead more specifically characterized as an energetic movement, the high-probability words could include 'sprints' or 'flies'. Of course, these are only two of the many possible relationships that could describe the words' meanings in relation to one another.

Hierarchical Bayesian program learning (HBPL) (Lake, Salakhutdinov, and Tenenbaum 2015) provides a useful framework for modeling story artifacts as factorized joint probability distributions. Let a story $S = w_1, w_2, \ldots, w_n$ be a sequence of n words w_i with $w_i \in \mathscr{W}$, the set of all possible words. Then a probabilistic approach to the problem of story creation imposes a joint distribution $p(S) = p(w_1, w_2, \ldots, w_n)$ over the set $\mathscr{S}$ of all possible stories S. Thus, creating a story means simply sampling from $p(S)$. Of course, for stories of any length, this distribution is likely to be intractable to compute, and thus typically some simplifying assumptions are made that allow the joint distribution to be factored in some way.

HBPL suggests that there are domain-specific factorizations that both simplify the computational demands of this generational approach and that exhibit explanatory power as well. Such factorization mirrors the human creative process both by providing focus and by reducing the large space of artifact creation to a series of smaller, tractable creative decisions.

For the domain of n-word microfiction stories, the most complete factorization comes from an application of the chain rule:

$$p(S) = p(w_1, w_2, \ldots, w_n) = p(w_{i_1})p(w_{i_2}|w_{i_1}) \ldots p(w_{i_n}|w_{i_{n-1}})$$

where $i_j \in [1, n]$ and $i_j \neq i_k$ unless $j = k$, so that this represents a general version of the chain rule that admits any possible permutation of word order dependency.

Thus, we have a formal model for hierarchy structure that allows any ordering of relationships between words and their corresponding positions in the final story. The factorization is hierarchical—each subdistribution beyond the first is conditioned on a word chosen from a subdistribution preceding it—and the formalism says nothing about how the probabilities for a given subdistribution are calculated, allowing for any and all interpretations of the relationships between the words in the story. If the structure and relationships imposed by a given factorization are "good", the result should be "good" stories. Similarly, a good factorization should be more tractable to compute than the whole joint probability.

In what follows, we refer to a particular factorization of the joint as a story *format* that represents one possible hierarchical structure. The relationships described by the conditional subdistributions of a given format determine what the format represents. The trivial format is one in which each subdistribution assigns all probability to a single word; sampling from that joint will always yield the same story artifact. A better format could, for example, capture linguistic relationships, assigning high probability to all words of a given part of speech. A further improved format could represent the semantic relationships between words and assign high probabilities to words that combine to form meaningful stories.

Computationally modeled formats with tractable subdistributions would be powerful tools to aid machine comprehension and generation of creative artifacts. Such formats could be manually constructed, as in the MICROS system, but automatically inferring the underlying semantic format of a human-written exemplar allows the system more flexibility and breadth of results. Indeed, we argue that such a system exhibits more creativity.

We have developed a generalized six-word story writing system that leverages the power of this model to create stories whose underlying structures mirror the linguistic coherence and semantic richness typified by human-written stories.

HIEROS

To test the efficacy of our model we developed HIEROS, a creative system that writes six-word stories using the inferred formats of human-written stories to guide the linguistic and semantic choices it makes when selecting words to create a novel artifact. The system operates in three main steps: inferring formats from exemplar stories, generating new stories based on those formats, and evaluating those stories by assigning each a quality score. Generated and scored stories are then refined by repeatedly mutating high-scoring stories via a modified generation step.

Inferring Story Formats

In order to model the meaning and quality of human-written stories for use in novel artifact creation, HIEROS infers formats from a list of exemplar stories that the system takes as input. In particular, the inferred formats contain linguistic (part-of-speech) information as well as an approximation of the semantic relationships typified by the words in the exemplar. The system first constructs a hierarchy of parts of speech and dependency relationships then fills in semantic information pertaining to those relationships.

The structure of the format is constructed using the Stanford Parser (De Marneffe, MacCartney, and Manning 2006) which statistically parses the exemplar and extracts both the parts of speech of each word and the dependencies between the words, forming a parse tree of dependency relations. Such relations reflect directed linguistic links between words in the artifact.

Given a six-word story exemplar $X = x_1, x_2, x_3, x_4, x_5, x_6$ with $x_i \in \mathscr{W}$, the parser constructs a directed acyclic graph $G = (V, E)$ with $V = X$. Each dependency in the parse tree is represented by a directed edge $(x_p, x_c) \in E$ where x_p is the parent word and x_c is the child word. One word x_r has no parent and serves as the root of the parse tree.

The dependencies represented by edges in the parse tree dictate the ordering of the words w_1 through w_6 in the format's factorization of the joint $p(S)$, such that

$$p(S) = p(w_r) \prod_E p(w_c|w_p).$$

Finally, $\psi(x_i)$, where ψ is a function that returns a word's part of speech, is recorded for each $i \in [1, 6]$, completing the linguistic hierarchy.

Automatically extracting semantic information from text is challenging. Although HIEROS' scope is limited to identifying relationships between just two words at a time, its method for constructing $p(S)$ must be able to model all possible semantic relationships between words that may exist in the formats inferred from exemplar inputs. To model such semantic relationships, HIEROS draws on the linguistic theory of distributional semantics, which hypothesizes that words that have similar meanings appear in similar contexts (Sahlgren 2008).

Specifically, HIEROS uses word2vec word embeddings (Mikolov et al. 2013) to model distributional semantic information. Being trained on a large corpus of data, in this case Wikipedia, the word embedding model represents each word's aggregate context as a many-dimensional vector. Thus, all the words in the corpus are mapped into one vector space that represents their relative contexts and therefore relative meanings.

HIEROS uses the difference between two words' word embedding vectors as a representation of the semantic relationship between those two words. The system traverses each edge in the exemplar's parse tree and calculates $\upsilon(x_c) - \upsilon(x_p)$, where υ is the word embedding function. It records the resulting *semantic vector* for use in identifying other words that share the semantic relationship that the edge represents. Thus, the system can be said to reverse engineer the semantic relationships between words in a human-written exemplar to use as a template for generating novel stories.

With these steps completed, HIEROS has constructed a format that represents the hierarchical ordering of the words in the exemplar, their parts of speech, and the semantic relationships between the dependent words. This corresponds to a factorization of $p(S)$ in which each subdistribution is conditioned on the word that precedes it in the hierarchy. The part-of-speech and semantic vector data recorded for each subdistribution will be used at generation time to assign high probabilities to words that fulfill the linguistic and semantic restrictions corresponding to those values. The format informs how the subdistributions should be constructed; the generator constructs and samples from them to create a story.

Figure 1 shows an example story (1) that is parsed to form a hierarchy of dependencies (2). That hierarchy is converted into a format (3) which can be visualized as a graph in which each node records an index into the story and a part of speech, and each edge stores the semantic vector calculated from the corresponding words in the exemplar. Note that the exemplar words themselves do not form part of the format.

The exemplar's parse tree dictates the format's factorization of the joint, in this case:

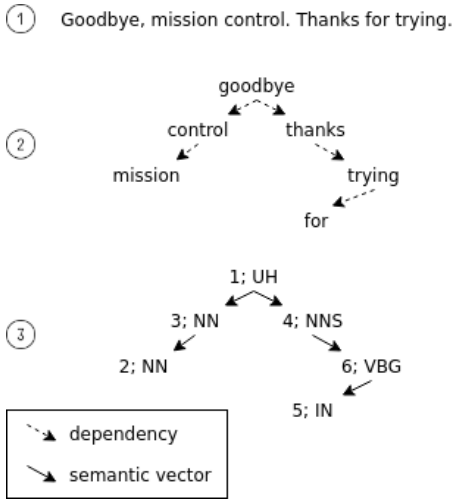

Figure 1: Example of format inference from an exemplar.

$$p(S) = p(w_1, w_2, w_3, w_4, w_5, w_6) =$$
$$p(w_1)p(w_3|w_1)p(w_4|w_1)p(w_2|w_3)p(w_6|w_4)p(w_5|w_6).$$

Recall that for each subdistribution $p(w_i|w_j)$—corresponding to the edge (x_j, x_i) in the hierarchy graph—the format stores an associated semantic vector $\upsilon(x_j) - \upsilon(x_i)$ and for each w_i the format stores a part of speech $\psi(x_i)$. These values are used to construct each subdistribution.

Generation

Our HBPL model of story creation views writing a six-word story as sampling from $p(S)$. To accomplish this, the generator selects one inferred format at random, calculates the probability distributions for each factor of the joint represented by that format, and samples them to yield a six-word story.

Sampling from Subdistributions HIEROS begins generation by constructing the first subdistribution $p(w_r)$, which represents the root of the format's hierarchy, and sampling from it. Because no word precedes the root, there is no semantic relationship to consider when assigning probabilities, and the subdistribution is populated by common words that match the given part of speech. To identify these words, HIEROS employs a list of the most common English words from the Corpus of Contemporary American English (Davies 2010).

To populate the root subdistribution, all words from this list that match the appropriate part of speech are collected and sorted from most common to least. Then the top 10% of the words, rounded down, are discarded in order to filter out bland or overly common words. The remaining words are each assigned equal probability to form the first subdistribution, which HIEROS samples to select the root word.

The format's hierarchy dictates an ordering for building the remaining subdistributions $p(w_c|w_p)$ and sampling w_c

for $c \in [1, 6] - r$. HIEROS constructs these subdistributions as follows.

The independence assumptions provided by our model's application of the chain rule reduce the large space of possible words into the tractable problem of identifying words of a certain part of speech that fulfill a given semantic relationship with a preceding word. Recall that the inferred format contains for each subdistribution a part of speech $\psi(x_c)$ and a vector $\upsilon(x_c) - \upsilon(x_p)$ representing the semantic relationship that should exist between w_p and the words with non-zero probability in $p(w_c|w_p)$.

The vector $\vec{v} = \upsilon(w_p) + \upsilon(x_p) - \upsilon(x_c)$ corresponds to a point in the word embedding space, such that the set

$$Y = \{y \in \mathscr{W} \mid \alpha > \|\vec{v} - \upsilon(y)\| \wedge \beta < \|\vec{v} - \upsilon(y)\|\}$$

contains words that are related to w_p according to the relationship dictated by the semantic vector. α and β are constants that bound the minimum and maximum distances of a word's embedding from $\vec{v}$ for it to be considered related to w_p in this way. $\beta > 0$ prevents words that are too closely related to w_p from populating the subdistribution, as these words are likely bland or uninteresting.

HIEROS constructs $Y' = \{y' \in Y \mid \psi(y') = \psi(x_c)\}$, resulting in the final list of words to which equal probabilities will be assigned to populate $p(w_c|w_p)$. HIEROS samples w_c from the resulting distribution.

Once all six subdistributions have been constructed and sampled, the system has sampled $S = w_1, w_2, w_3, w_4, w_5, w_6$ from $p(S)$, completing one round of generation.

Mutating via Resampling As refining HIEROS' generated stories involves mutating a generated story, the generator is designed to efficiently resample $p(S)$ by selecting a word w_m and drawing a new word $w'_m \neq w_m$ from the same subdistribution from which it was originally drawn.

Due to the hierarchical nature of the format, this reselection may necessitate cascading changes to other words in the story. If the selected word has dependencies below it in the format's hierarchy, those subdistributions will be constructed anew, conditioned on the newly selected preceding word, and sampled to select new words. If dependencies are found below those words, those subdistributions will be reconstructed and resampled, and so on. Thus, by changing one word in the story, the mutation process may result in a new story that differs from the original by more than one word.

Evaluation & Refinement

HIEROS improves the quality of its stories by evaluating and refining them. It scores its generated stories and mutates them to observe whether resampled stories score higher, continuing until no higher scoring stories are generated.

Scoring Assigning scores to creative artifacts that reflect their quality is a key challenge for any computational system that relies on a generation-evaluation loop. HIEROS employs the same skip-thought scoring method as MICROS. Skip-thought vectors (Kiros et al. 2015) are similar to word2vec word embeddings in that they encode natural language strings as vectors in a high-dimensional semantic space, but whereas word2vec maps words to vectors, the skip-thoughts model maps sentences to vectors.

Skip-thought vectors can be used to score a story S_s by measuring its similarity to two high-quality stories S_g, S_h and its dissimilarity from a poor-quality story S_b, where $S_i \in \mathscr{S}$.

Let $\vec{a_1} = \tau(S_g) - \tau(S_b)$ and $\vec{a_2} = \tau(S_h) - \tau(S_b)$, where τ is the story embedding function. Let $\vec{n} = \vec{a_1} \times \vec{a_2}$, and let $\phi_{\vec{n}}$ be the function that projects a vector onto $\vec{n}$. Let $\vec{s} = \tau(S_s)$. Then the score for S_s is $\|\vec{s} - \phi_{\vec{n}}(\vec{s})\|$, or the magnitude of $\vec{s}$ when it is projected onto the plane defined by $\vec{a_1}$ and $\vec{a_2}$ whose origin is at $\tau(S_b)$. Thus, poor-quality stories have vector representations that project closer to the origin, and the vectors for higher-quality stories are projected further from it.

Because the scoring plane is described by vectors from S_b to S_g and S_h, we refer to the high-quality stories as "axes".

Choosing which stories to use as axes is a critical consideration for this scoring method. HIEROS leverages its list of exemplar inputs to select high-quality axis stories for scoring.

Ideally, the two axes are distinct in order to capture as much information in the score as possible. Furthermore, the axes should ideally not include the exemplar story that inspired the generated story to be scored; otherwise high scores will be assigned to generated stories that are very similar to their exemplars. However, if the axes are too different from the story to be scored, the score will be less accurate. To balance these considerations, HIEROS prefers to select axes that are distinct from the exemplar story but that still share similar parts of speech with that story.

Let $P(X)$ represent the concatenation of exemplar X's parts of speech: $P(X) = \psi(x_1)\|\psi(x_2)\|...\|\psi(x_6)$. Let $C = \{(X_k, X_j) \mid P(X_k) = P(X_j)\}$. For each exemplar X_i, HIEROS chooses at random X_j such that $(X_i, X_j) \in C$, with $i \neq j$ if possible. X_i and X_j then serve as scoring axes S_g and S_h for all stories generated with the format inferred from X_i. If $X_i \neq X_j$, we refer to these axes as "diverse", otherwise we refer to them as "single" axes.

We experimented with using different axis configurations to score HIEROS' stories. We discuss the trade-offs between different axis selection methods in a later section but note here that we chose to use diverse axes.

Refinement Process HIEROS' refiner organizes generated stories by root word, maintaining a priority queue containing the highest scoring stories with that root that it has generated so far. These queues are initially seeded by generating one story for each possible root word (i.e. each word with non-zero probability in the format's first subdistribution) instead of selecting a root randomly. With the priority queues thus populated, refinement then proceeds in steps.

At each refinement step, the highest scoring story in each queue is popped and mutated to generate a specified number of children. Each of those children is scored and placed into the priority queue corresponding to its root (which may differ from its parent's due to mutation). Each queue also

remembers the highest scoring stories it has seen across all refinement steps. After each step, if its highest-scoring story has not changed, a counter reflecting that queue's "staleness" is incremented. Once that counter reaches a specified number, the queue is considered stale and no more stories are popped from it. When all queues are stale the refinement process terminates, and the two highest-scoring stories from each queue are collected to form the output of the refiner.

Due to the limitations of skip-thought scoring, some stories receive higher scores without being of higher general quality than other stories. This bias appears to be linked to certain words in a story. By maintaining several queues and taking the top stories from each, the system avoids biasing its output toward many similar stories that feature such words.

After refinement, the final step in generating stories for a given format is to capitalize and punctuate them according to the format's exemplar.

To take advantage of the variety of formats that may exist among the exemplar stories, HIEROS repeats this same generation and refinement process for several randomly selected formats. After refined stories have been generated for each format, the combined results are sorted by score, and a subset of the highest scoring stories is returned as the creative output of the system.

This subset includes a number of high-scoring generated stories, excluding a number of the highest-scoring of those stories, with both the specific numbers of inclusion and exclusion specified by parameters. Due to the scorer's tendency to assign high scores to stories that more closely resemble the exemplar, the most interesting potential output may not include those stories with the highest scores.

Results

HIEROS' results demonstrate that our formulation of story formats is a useful model to guide the automatic creation of six-word stories. Its ability to model the underlying structure of human-written stories results in broad variety among its created artifacts.

Coherence and impact are two complementary qualities by which six-word stories may be judged. Coherence refers to whether a story is understandable and makes sense, while impact refers to whether a story succeeds in eliciting an emotional reaction from the reader. These qualities parallel the latent structures that our system attempts to infer from exemplar stories, namely linguistic structure and semantic structure. A format that accurately models the former should generate coherent stories, while a format that models the latter has a higher chance of generating impactful ones.

Input & Parameters

We scraped exemplar six-word stories from Reddit[1] and Twitter[2][3]. We removed any stories that featured nonstandard spelling or symbols as well as others we deemed unsuitable for format inference, including stories that featured acronyms or pop culture references. This left us with 1,481 exemplars from which HIEROS inferred formats. HIEROS selects formats that have diverse scoring axes as candidates for generation first, falling back to those with single axes if it finds an insufficient number that are diverse. 261 of the inferred formats could be clustered by part of speech with one or more other formats, giving them two distinct scoring axes. These formats were used for story creation.

The parameters HIEROS used for story creation are as follows. Four children were created every time a story was popped from the top of its priority queue in the refiner, and the queue for a given root was considered stale if three rounds had passed without it seeing a new highest scoring story. Each time HIEROS ran, it generated stories for 30 formats chosen at random and selected as its output portion the 85th to 90th percentile of highest scoring stories. This resulted in the system outputting approximately ten stories each execution.

Survey

We conducted a survey to evaluate HIEROS' results and the efficacy of its scorer using the output of ten executions—for a total of 100 stories—plus the 15 lowest scoring stories that the refiner saw over one execution. The latter group of stories would never be output from the system, but we included it to serve as an indication of how well the refiner is able to distinguish low-quality stories.

The survey briefly introduced six-word stories as a type of creative artifact, defined coherence and impact, and presented 15 randomly chosen stories for the respondent to evaluate. For each story, we asked the respondent to rate the degree to which they agreed with the following two statements on a seven-point Lickert scale: "This story is coherent (understandable, correct English)." and "This story is impactful (meaningful, funny, sad, etc.).".

We distributed the survey via social media and received 124 responses, including partial responses in which fewer than 15 stories were rated. We did not collect demographic data as part of the survey, but the majority of the audience to which the survey was presented were native English speakers. Each story was rated by an average of 14 respondents.

The results of the survey demonstrate that a handful of HIEROS' stories achieve coherence, such as "To him, 'endlessly' meant 'twenty decades'." and "Joy is a path to childhood.", however the majority do not. Very few of its stories are both coherent and impactful, however some do manage to achieve this more elusive effect, such as "Diamond ring. Glassy diamond. Costliest engagement.". Interestingly, while most stories were rated as having lower impact than coherence, some were found to be impactful despite being somewhat incoherent. One such example, "Find, yours cowardice is his strength.", seems to evoke a response through the emotionally charged juxtaposition of "[your] cowardice" and "his strength" despite not telling an understandable story.

We can characterize the accuracy of the skip-thought scorer by comparing the ratings respondents gave to the top 15 highest skip-thought-scored stories compared to the 15 lowest scored stories. Figure 2 shows boxplots of the ag-

[1]https://www.reddit.com/r/sixwordstories/top/?t=all

[2]https://twitter.com/sixwordstories

[3]https://twitter.com/ernest6words

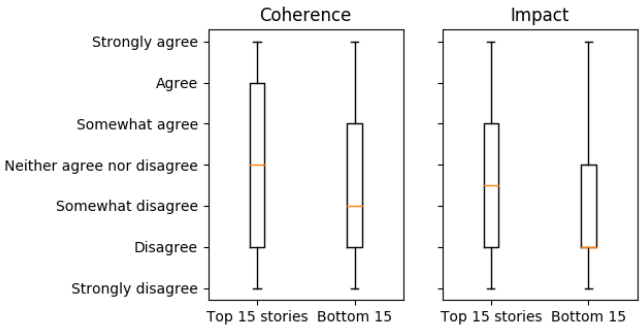

Figure 2: Boxplot comparison of human-rated coherence and impact for top and bottom skip-thought-scored stories. Differences in both characteristics are statistically significant.

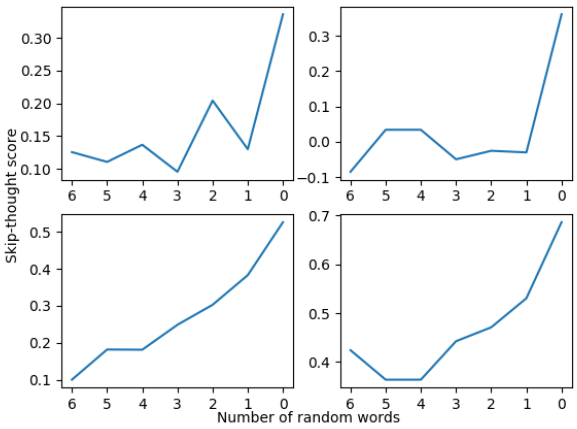

Figure 3: Line graphs of the skip-thought scores of four randomly selected HIEROS-generated stories with decreasing numbers of story words replaced with random words. Six random words is a completely random story, five random words is the generated story with every word except for the first replaced, four random words sees all but the first two words replaced, and so on to zero random words which is the original generated story.

gregate coherence and impact ratings of these two groups of stories. The respondents' ratings of the top 15 stories' coherence were significantly higher than their ratings for the bottom 15, $t(208) = 3.01, p = 0.003$. The respondents' impact ratings were also significantly higher for the top 15 stories than for the bottom 15, $t(208) = 4.30, p < 0.001$, using an alpha of 0.01 for both statistical tests.

This demonstrates that the skip-thought scorer successfully assigns lower scores to low-quality stories. However, the scorer also assigns high scores to many stories to which respondents assign low ratings. Thus the scorer, and by extension the refiner, is limited in its ability to truly distinguish high-quality six-word stories.

Discussion

HIEROS' results show that the system, while imperfect, is capable of writing interesting stories. Furthermore, because the system takes as input any set of six-word stories, its output is as varied as its input. This is demonstrated by HIEROS-created stories such as "No-one persists sympathetic. Dislike is unsympathetic.", "His dirtiest bush is seldom artificial.", "Adverse anger yells. Undesirable euphoria grieves.", "A creamy maroon color turned currant.", and "Disgust fixes each tenderness within misadventure.". This ability to create a variety of occasionally-interesting stories indicates that HIEROS' inferred formats do capture some degree of the exemplars' structure and semantic meaning.

HIEROS' refiner uses its skip-thought scorer to explore the myriad of stories represented by $p(S)$. The interplay between the quality of stories that are assigned high probabilities in that joint distribution and the stories to which the scorer assigns high scores determines the quality of the system's final output. Axis selection is critical to the accuracy of the skip-thought scorer. We experimented with two main approaches to selecting axes: choosing X_j at random such that $(X_i, X_j) \in C \wedge i \neq j$ and choosing X_j, X_k at random such that $(X_i, X_j) \in C \wedge (X_i, X_k) \in C \wedge i \neq j \neq k \neq i$.

When the exemplar X_i is used as a scoring axis, the scorer assigns high scores to stories that are more coherent but also more similar to the exemplar. When the exemplar is not included among the axes, we observe that high-scoring stories are more varied but of lower quality overall. In order for the score to better reflect story quality, we decided that it was permissible to bias the scorer somewhat toward stories that were similar to the exemplar.

To evaluate whether the skip-thought scoring method achieved its goal of reflecting story quality, we performed an experiment in which we scored HIEROS-generated stories as progressively more of their words were replaced with random words. If the scorer is able to accurately measure quality, then a story should score lower the more random words are present in it. Figure 3 shows the results of conducting this experiment on four randomly selected HIEROS-generated stories. These results demonstrate a general trend that less random stories correlate to higher scores.

Despite this demonstration that HIEROS' scoring method can distinguish between random and non-random stories to some degree, it is clear that the scorer is primarily to blame for the low quality of the system's stories overall. As evidenced by the best of its output, HIEROS' generation and mutation systems are capable of writing quality stories. However, the scorer is not able to consistently identify those stories. A more accurate scorer would be more effective in directing HIEROS' story generation and would improve the system's results.

Similarly, the word2vec word embedding model that HIEROS uses could be improved or replaced in order to improve the quality of the system's output. HIEROS' word2vec model was trained on a Wikipedia corpus that, while large, does not include more poetic contexts for the words it contains, restricting the breadth of the model's rep-

resentations of those words meanings. Training word2vec on a corpus of poetry or other literary works is thus a potential avenue for improving HIEROS. Alternatively, replacing word2vec with an improved model of semantic meaning could help prevent HIEROS' word selection from tending toward synonyms of exemplar words.

Finally, we note that the quality of a HIEROS-generated story is unlikely to surpass that of its exemplar story. Although the exemplar stories scraped from Reddit and Twitter represent relatively high-quality amateur stories, using exemplars written by more skilled poets would raise the ceiling on the quality of HIEROS' stories.

Conclusion

Framing a creative artifact's structure as a hierarchical factorization of a joint probability over potential elements allows a computational system to place useful restrictions on the elements it selects to generate novel artifacts. This model provides a tractably computable means for modeling the cohesion and richness of human-created artifacts, which is useful for imparting quality and generalizability to a creative system.

We have presented a model of hierarchical artifact structure, examined how it can be applied to the domain of six-word stories, and demonstrated a method for inferring structure from exemplar stories. Our creative system HIEROS produces a wide variety of output using this method, some of which achieves impactfulness, providing an argument for the utility of this model as a guide for creative generation.

References

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-FACE Poetry Generation. In *Proceedings of the 3rd International Conference on Computational Creativity*, 95–102.

Davies, M. 2010. The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25:447–464.

De Marneffe, M.-C.; MacCartney, B.; and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation*, volume 6, 449–454.

Fillmore, C. J., and Baker, C. F. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, volume 6.

Fillmore, C. J.; Wooters, C.; and Baker, C. F. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, 3–26.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, 3294–3302.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350:1332–1338.

León, C., and Gervás, P. 2014. Creativity in story generation from the ground up: Non-deterministic simulation driven by narrative. In *Proceedings of the 5th International Conference on Computational Creativity*, 201–210.

Li, C.; Wu, Y.; and Lan, M. 2018. Inference on syntactic and semantic structures for machine comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv* abs/1301.3781.

Pérez y Pérez, R., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13:119–139.

Riedl, M. O., and Young, R. M. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24:303–323.

Sahlgren, M. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20:33–53.

Spendlove, B.; Zabriskie, N.; and Ventura, D. 2018. An HBPL-based approach to the creation of six-word stories. In *Proceedings of the 9th International Conference on Computational Creativity*, 161–168.

Toivanen, J.; Toivonen, H.; Valitutti, A.; Gross, O.; et al. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the 3rd International Conference on Computational Creativity*, 175–179.