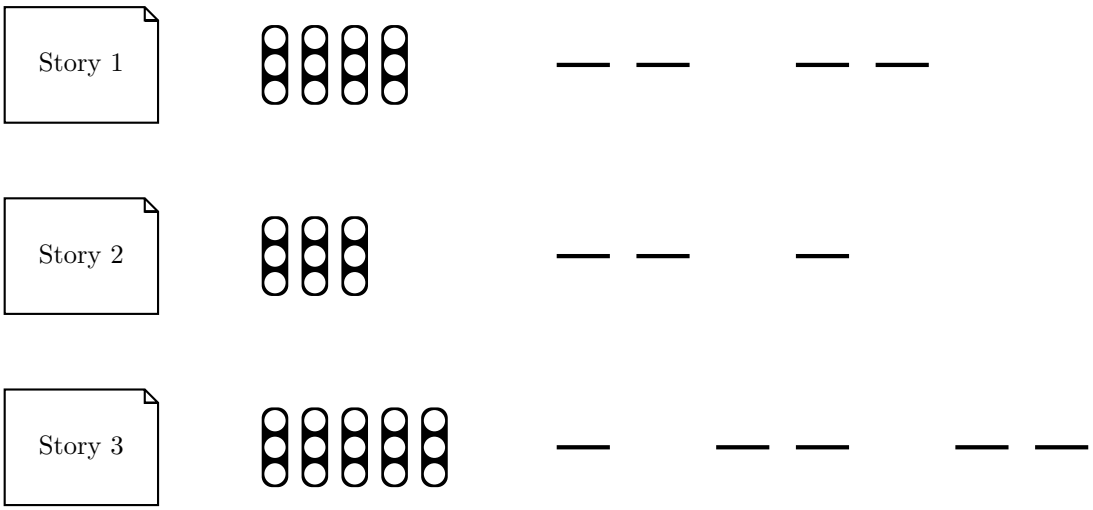

Figure 1: Caption

Storylines

November 16, 2020

1 Introduction

Modern generative models of text are extremely flexible and accurate, but may become incoherent when generating long texts due to a limited context window [3**?**]. Additionally, the popular approach to language modeling relies on opaque and difficult to control models. These two flaws limit the applicability of modern models to story generation, where coherent long-form generations are required and controllability is necessary for quality control. In such scenarios, a model with explicit hierarchical structure is much more suitable, as hierarchical structure separates the low-level process of generating words and sentences from story structure.

Our goal is to design structured generative story models by modeling storylines. Prior work notes that including structure in the generative model improves coherence [4, 8, 11, 19**?**]. However, prior investigations into narrative structure use hand-crafted representations, such as entity coreference or keywords. We explore a different definition of what constitutes a storyline: We hypothesize that storylines are composed of groups of sentences, or segments, that function similar to biological sequence motifs.

Biological sequence motifs are recurring patterns that have biological significance [6]. For example, a motif in DNA may determine how protein interactions take place and play a role in regulating gene expression. Analogously, we posit that storyline motifs determine how a reader interprets a story. An example of a storyline motif is a sequence of sentences that finishes introducing characters and setting the story. This motif would signal the end of introduction, preparing the reader for the start of the rising action.

Ideally, we could obtain manually labeled storyline motifs from humans, as we have defined motifs to be recurring patterns with a similar effect on readers. However, this process would be very expensive. It would require both designing an experimental procedure for labeling motifs and also obtaining annotations from humans, both of which are costly. Thankfully, biological sequence alignment algorithms offer a solution. Biological sequence alignment algorithms find motifs without directly modeling downstream effects, using only a measure of similarity between elements of the sequences. Given the similarity measure, alignment algorithms do not require further learning.

Typically, these similarity measures are derived from manually-specified rules specific to biological applications. Our hypothesis, that similar storylines segments have a similar effect on readers, suggests that the similarity measure for storylines should be semantic. We therefore turn to large pre-trained sentence representations [17], as they have been shown to perform very well on semantic similarity tasks such as natural language inference and entailment.

We perform an exploratory analysis of storyline induction by utilizing biological sequence alignment algorithms as well as large pretrained sentence representations. We show that alignment algorithms built upon measures of semantic similarity are able to extract common structure from stories.

2 Related Work

Automatic story generation is a difficult task. The best-performing language models are extremely flexible, but not specialized for generating coherent stories. Additionally, highly flexible black-box models are difficult to control, as they were not designed with intervention in mind. Prior work addresses coherence by using incorporating hand-designed proxies of storylines, such as keywords [11, 19], information from classical NLP pipelines [8], or visual information [4]. Although effective at improving the generative model, approaches that require specifying storyline representations are unlikely to scale to many use-cases and diverse domains.

A closely related work, by Papalampidi et al. [14], formulated a latent variable model for narrative structure in screenplays. Their work uses a pre-trained model trained on manually labelled narrative

structure, and found it transferred well to their new setting. They fine-tune the pretrained model on their dataset as part of a summarization model, which creates a short summary given a single source screenplay. Our approach is more general as it does not assume a model with knowledge of narrative structure, just semantic similarity. Additionally, we generalize to the case of multiple sources in order to extract robust motifs.

Another line of work seeks to increase the controllability of black-box generative models without adding structure. One approach is to train models for story infilling, which propose sentences to insert in the middle of a story [10], and another is to learn post-hoc controllability [5]. Both of these approaches suffer from a lack of interpretability, which complicates interaction with users. A hierarchical generation process allows for a user interface to be an integral part of the model.

Work in structured molecule generation also benefits from a hierarchical generation process. The structured generative model in Jin et al. [13] utilize graph motifs were useful for improving a generative model of molecules. Although the motifs in molecules are graph-based, and therefore both easier to identify and structurally different than the sequence motifs we consider, the improvements to the generative model are significant.

A recent application of biological sequence algorithms by Alayrac et al. [1] applies techniques from alignment to the problem of aligning instructional videos to textual narrations. Analogous to storyline induction from multiple stories, they aim to jointly align multiple narratives to video in order to discover actions sequences. Interestingly, they find that classical approaches to sequence alignment are outperformed by modern optimization techniques. However, storylines are much more complex than action sequences from instructional videos, as the intentions behind stories vary greatly.

3 Background: Pairwise Sequence Alignment

Pairwise sequence alignment is used in bioinformatics to measure the similarity between two sequences. Alignment algorithms extend a measure of similarity from operating on individual elements to operating on entire sequences. This is accomplished by establishing an element-to-element correspondence between the elements of a pair of sequences such that ordering is preserved. This correspondence is obtained by inserting gaps into both sequences. In our case, instead of biological sequences, we are interested in aligning stories, i.e. sequences of sentences.

Pairwise sequence alignment finds the minimum cost alignment between sequences of sentences $\mathbf{x}_1$ and $\mathbf{x}_2 \in \Sigma^*$, where sentences are represented by a vector in $\mathbb{R}^n$. The vocabulary Σ contains a gap symbol '-', such that $\Sigma = \mathbb{R}^n \cup \{-\}$. We refer to elements of Σ, which may be a sentence or gap, as tokens.

Given the two stories $\mathbf{x}_1$ and $\mathbf{x}_2$, a pairwise alignment is given by $\mathbf{x}_1^\pi, \mathbf{x}_2^\pi$, where $\mathbf{x}_1^\pi$ and $\mathbf{x}_2^\pi$ are 1) of the same length L and 2) ignoring gaps, are identical to $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. In order to align these stories, pairwise alignment extends edit distance with a distance measure $d : \Sigma \times \Sigma \to \mathbb{R}$ that operates on pairs of tokens. The best alignment is given by the following:

$$\underset{\pi}{\operatorname{argmin}} \sum_{(i,j)\in\pi} d(x_{1i}^\pi, x_{2j}^\pi), \tag{1}$$

where π denotes where gaps are inserted in each sequence. Equation 1 can be solved exactly via dynamic programming. We refer to $M = \begin{pmatrix} \mathbf{x}_1^\pi \\ \mathbf{x}_2^\pi \end{pmatrix}$ as the alignment.

When choosing a distance measure d, a key component is how the measure deals with gaps. There are two approaches: assume that 1) gaps correspond to insertion or deletions and 2) gaps correspond to expansions or compressions. Classical algorithms from biology, such as Needleman-Wunsch, model insertions and deletions, as random mutations result in completely new elements in DNA. Other algorithms such as Dynamic Time Warping (DTW) model expansions and compressions, since warping may occur due to various reasons such as sampling at different frequencies. With stories, an example of expansion would be to split a long sentence into shorter sentences. Although it is possible to model both of these phenomena in the distance measure, we assume that adding a sentence to a story is closer to a mutation than an expansion. A new sentence is likely to add new information and therefore be semantically different from preceding sentences. We therefore only model insertion and deletion in the distance measure.

4 Problem Setup: Multiple Sequence Alignment

In order to extract storylines, we would like to find patterns that are robust across multiple stories. Unfortunately, methods for solving pairwise alignment are not directly applicable, since they only consider pairs of stories. We instead turn to the problem of multiple sequence alignment, which generalizes alignment from pairs of stories to sets of stories.

Given a set of K sequences, a multiple alignment is a matrix $M \in \Sigma^{K \times L}$, given by:

$$M = \begin{pmatrix} \mathbf{x}_1^\pi \\ \vdots \\ \mathbf{x}_K^\pi \end{pmatrix}, \tag{2}$$

where each $\mathbf{x}_i^\pi$ is obtained by inserting gaps into the corresponding $\mathbf{x}_i$ to ensure that each $\mathbf{x}_i^\pi$ is of length L, as in pairwise alignment.

There are two common measures of quality for a multiple alignment: 1) the sum of pairs (SP) score and 2) the Steiner distance. All methods use the distance measure $d : \Sigma \times \Sigma \to \mathbb{R}$ from pairwise alignment. The SP score is given by

$$\mathrm{SP}(M) = \sum_{l=1}^{L} \sum_{i=1}^{K} \sum_{j=i+1}^{K} d(M_{il}, M_{jl}), \tag{3}$$

obtained by summing the pairwise distances between elements of M in the same column. The Steiner distance can be computed without explicitly constructing a multiple alignment. Rather than operating on a multiple alignment, the Steiner distance instead measures the distance from an average sequence to each of the sequences $\mathbf{x}_k$. In order to define the Steiner distance, we need a measure of distance between sequences. Let $D : \Sigma^* \times \Sigma^* \to \mathbb{R}$ be such a measure, obtained via a pairwise alignment using the token distance d. The Steiner distance is then given by

$$\min_{\mathbf{z}} \mathrm{SD}(\mathcal{X}, \mathbf{z}) = \min_{\mathbf{z}} \sum_{i=1}^{K} D(\mathbf{x}_i, \mathbf{z}), \tag{4}$$

where the optimal $\mathbf{z}$ is the average sequence, also referred to as the Steiner sequence. Given a set of stories, their average sequence $\mathbf{z}$ has a neat intepretation as a common storyline.

We utilize algorithms that optimize either the SP score and the Steiner distance in order to perform multiple alignment.

5 Methods

As optimizing both the SP score and Steiner distance are NP-hard, we resort to heuristic optimization methods. We use a combination of three approaches: a progressive alignment method inspired by a classical algorithm from computation biology, an iterative averaging method similar to a popular algorithm from time-series analysis, and a greedy hill climbing algorithm that operates in the space of Steiner sequences.

5.1 Progressive Alignment

Inspired by the progressive alignment algorithm of Feng and Doolittle [9], we take a greedy approach to approximating the multiple sequence alignment problem. Progressive alignments aim to optimize the SP score, and have been found to perform well despite their simplicity.

Given a set of sequences $\mathcal{X}$ and an ordering σ, we progressively align the next sequence in the ordering to the already aligned sequences. Once a set of sequences are aligned, their columns of the alignment are frozen; elements of new sequences must align to a whole column from the existing alignment or a gap. This is referred to as the 'once a gap, always a gap' property [9].

In order to align a sequence to an existing alignment, we lift the definition of the token distance d to compare an element to a column of an alignment such that $d^+ : \Sigma^* \times \Sigma \to \mathbb{R}$, as follows:

$$d^+(\mathbf{y}, x) = \sum_{j=1}^{|\mathbf{y}|} d(y_j, x). \tag{5}$$

We can then extend pairwise global alignment with d^+, allowing us to align the columns of an alignment to a token. The full algorithm is given in Algorithm 1.

Algorithm 1 Progressive Alignment

Given: A set of sequences $\mathcal{X} = \{\mathbf{x}_i\}_i$, ordering σ, and extended distance measure d^+
Initialize alignment M to $\mathbf{x}_{\sigma(1)}$
for all sequences $\mathbf{x}_{\sigma(i)}$ in order σ **do**
    Update $M = \textsc{PairwiseAlign}(M, \mathbf{x}_{\sigma(i)}, d^+)$
return M

5.2 An Iterative Averaging Algorithm

Next, we propose an iterative averaging (IA) algorithm which directly optimizes the Steiner distance. The IA algorithm is based on the DTW Barycenter Averaging (DBA) algorithm [16], which iteratively computes pairwise alignments between a mean sequence $\mathbf{z}$ and a set of sequences $\mathcal{X}$ then uses those alignments to recompute the mean sequence. As the DBA algorithm was designed to model expansion and compression, we adapt it for insertion and deletion.

The IA algorithm proceeds in two steps: We first construct a multiple alignment from the pairwise alignments, then for each column of the multiple alignment compute the token that minimizes the distance to each element within that column. As we model insertions and deletions, there is ambiguity when mapping the pairwise alignments to a joint multiple alignment. Namely, there are multiple ways of aligning tokens from $\mathcal{X}$ that are all aligned to gaps in $\mathbf{z}$. We resolve this ambiguity heuristically by inserting additional gaps into the $\mathbf{x}_i^\pi$ obtained by pairwise alignment with $\mathbf{z}$ so that all tokens aligned to elements in $\mathbf{z}$ are aligned. The algorithm is detailed in Algorithm 2.

Algorithm 2 Iterative Averaging Alignment

Given: A set of sequences $\mathcal{X} = \{\mathbf{x}_i\}_i$ and distance measure d
Given: Initial mean string $\mathbf{z}$
function $\textsc{IterativeAverage}(\mathbf{z}, \mathcal{X}, d)$
    for all iterations **do**
        for all sequences $\mathbf{x}_i \in \mathcal{X}$ **do**
            Compute pairwise alignments $M^i = \textsc{PairwiseAlign}(\mathbf{x}_i, \mathbf{z}, d) \in \Sigma^{2\times L_i}$
        Compute multiple alignment $M = \textsc{Stack}(M^1, \ldots, M^K)$
        Update $\mathbf{z} = \textsc{Average}(M, d)$
    return M
function $\textsc{Stack}(M^1, \ldots, M^K)$
    for all indices j of $\mathbf{z}$ **do**
        for all sequences $\mathbf{x}_i$ **do**
            Set $g_{ij} = \#$ of elements of $\mathbf{x}_i$ aligned to z_j or a gap between z_j and z_{j+1}
        Set $g_j = \max_i g_{ij}$
        Set $\tilde{M}_i$ by inserting gaps into each M^i until there are $g_j - 1$ gaps after each z_j
    return $\begin{pmatrix} \tilde{M}^1 \\ \vdots \\ \tilde{M}^K \end{pmatrix}$
function $\textsc{Average}(M, d)$
    Initialize $\mathbf{z} \in \Sigma^L$
    for all columns l in M **do**
        Set proposal z' to the average non-gap representation of column l
        Compute costs $c(z') = \sum_i d(M_{il}, z')$ and $c(-) = \sum_i d(M_{il}, -)$
        if $c(z) > c(-)$ **then**
            Set $z_l = -$
        else
            Set $z_l = z'$
    return $\mathbf{z}$

5.3 Hill Climbing Algorithm

As the previous iterative algorithm used a heuristic in the averaging step that was computationally cheap but not guaranteed to improve the Steiner distance, we also consider a greedy hill climbing (HC) algorithm that is more expensive but only improves the objective.

The hill climbing algorithm proceeds as follows: Given an initial mean sequence $\mathbf{z}$, we first compute the pairwise alignments from $\mathbf{z}$ to each sequence in $\mathcal{X}$. We obtain an initial proposal sequence by averaging the token from each $\mathbf{x}_i$ aligned to each element of $\mathbf{z}$, rather than a gap. We then compute all one-step deviations from this proposal sequence, then find the proposal with the lowest Steiner distance to $\mathcal{X}$ for use as the mean sequence in the next iteration. One-step deviations are obtained by adding or deleting one element of $\mathbf{z}$. Candidates for addition are obtained by considering the elements of sequences in $\mathcal{X}$ that are aligned to gaps in-between elements of $\mathbf{z}$. The full algorithm is given below in Algorithm 3.

Algorithm 3 Hill Climbing Alignment

Given: A set of sequences $\mathcal{X} = \{\mathbf{x}_i\}_i$, and distance measure d
Initialize mean string $\mathbf{z}$
function HillClimb($\mathcal{X}, \mathbf{z}, d$)
  for all iterations **do**
    for all sequences $\mathbf{x}_i \in \mathcal{X}$ **do**
      Compute pairwise alignments $M^i = \text{PairwiseAlign}(\mathbf{x}_i, \mathbf{z}, d)$
    Compute proposal $\mathbf{z}' = \text{AverageAligned}(M^1, \ldots, M^K)$
    Initialize list $Z = [\mathbf{z}']$
    for all indices $t \in [|\mathbf{z}|]$ **do**
      for all alignments M^i **do**
        Store tokens in $\mathbf{x}_i$ aligned to gaps between z_{t-1} and z_t in g_i^t
      for all elements $\mathbf{y}$ of the cartesian product of $g_1^t \times \cdots \times g_K^t$ **do**
        Set m to the average representation of $\mathbf{y}$
        Append addition proposal $[\mathbf{z}_{1:t}, m, \mathbf{z}_{t:|\mathbf{z}|}]$ to Z
      Append deletion proposal $[\mathbf{z}_{1:t}, \mathbf{z}_{t+1:|\mathbf{z}|}]$ to Z
    if no proposals improve the Steiner Distance **then return z**
    Set $\mathbf{z}$ to the proposal in Z with the lowest Steiner distance
  return z
function AverageAligned($M^1, \ldots, M^K$)
  for all non-gap elements z_t in the alignments M_i **do**
    Set z_t to the average of all aligned tokens from each M_i
  return z

6 Experiments

We evaluate our MSA approaches on the WritingPrompts dataset [7], a dataset of 300K human-written short stories obatined from the WritingPrompts subreddit. Each story consists of a pair of a writing prompt and the story itself. In our experiments, we use only the story.

We perform MSA on story sets of size 5. As it would be intractable to run MSA on all story sets of size 5, we instead use the following procedure to obtain the sets $\mathcal{X}$:

1. For every story in the WritingPrompts dataset, we project each sentence using SBERT [17] into $\mathbb{R}^n$.

2. We compute the bigram bag-of-sentence (BoS) representations of stories by concatenating the SBERT representations of consecutive sentences and averaging over time.

3. We use each of the first 50k stories from the WritingPrompts dataset as centroids and find the 128 nearest neighbours for each centroid in bigram BoS space.

4. For each centroid, we then find the 4 closest stories from its 128 neighbours under the path-length normalized pairwise alignment distance. We avoid selecting duplicates by discarding stories with matching 10-grams.

5. Select 50 centroids (and their closest stories) based on the sum difference from the centroid to the closest stories.

For the token distance measure $d(x, y)$, we use

$$d(x,y) = \begin{cases} 0 & x = - \wedge y = - \\ \delta_x & x = - \wedge y \neq - \\ \delta_y & x \neq - \wedge y = - \\ \|x - y\|_2^2 & \text{otherwise,} \end{cases} \tag{6}$$

where δ_x, δ_y are gap penalties.

In choosing clusters, as well as our experiments with progressive alignment, we set $\delta_x = \delta_z \in \{125, 150\}$. For the Steiner distance, we set $\delta_x \in \{60, 75\}$ and $\delta_y \in \{180, 225, 300, 375\}$. We chose these gap penalties empirically based on preliminary analysis of the SBERT nearest neighbours of sentences as well as the resulting multiple alignments.

We compare multiple alignments based on the SP score and Steiner distance, and examine the output of each MSA algorithm by qualitatively evaluating the semantic closeness of alignments. For computing the MSA, we use the progressive alignment, iterative averaging, and hill climbing algorithm. For the IA algorithm, we initialize the mean sequence with the longest sequence $\mathbf{x}_i \in \mathcal{X}$. For the hill climbing algorithm, we initialize the mean sequence with two different configurations: either the mean sequence obtained from the progressive alignment or the IA algorithm.

7 Results

We find that each of the MSA algorithms perform better on the objective they optimize, as seen in Table 1. The progressive alignment performs well on the SP score, while the IA and hill climbing (initialized with the IA mean sequence) approaches perform well on the Steiner distance. Initializing the hill climbing approach with the mean sequence obtained from the progressive alignment results in an alignment that has a lower Steiner distance than progressive, but the SP score increases.

Algorithm	δ_x	δ_y	**SP Score**	**Steiner Distance**
Progressive	125	125	2,494,153.00	877,445.99
Hill Climbing (Progressive)	60	300	2,514,774.25	774,525.85
Iterative Averaging	60	300	2,643,249.25	716,350.33
Hill Climbing (IA)	60	300	2,647,854.25	709,076.44

Table 1: The SP scores and Steiner distances for each of the MSA algorithms. The progressive algorithm optimizes the SP score and achieves the lowest. The hill climbing algorithm initialized with the mean sequence obtained from IA obtains the lowest Steiner distance. We see the algorithms get stuck in local optima, as the hill climbing algorithm initialize with the progressive mean sequence obtains a much higher Steiner distance than if it had been initialized with IA.

We compare the method with the best SP score, the progressive alignment, with the best method in terms of Steiner distance, the hill climbing algorithm initialized with the output of iterative averaging (IA). The alignments obtained from the progressive alignments are longer and contain more gaps than the hill climbing (IA) alignments. This is shown in Figure 2, where the total number of columns is much larger for progressive than HC(IA), and the column densities are lower for progressive as well. We find that the hill climbing algorithm aligns sentences that are not semantically similar, and therefore focus on analyzing the progressive alignments for evidence of storylines.

We observe that stories with very short sentences result in qualitatively better alignments. This is corroborated by Figure 3, where the average distance from short sentences to their nearest neighbours is smaller than that of longer sentences. The shorter sentences may be better semantic matches, or there may be out-of-domain issues with the sentence representations from SBERT. The datasets SBERT was trained on were obtained from image captioning and other sources that may not transfer well to the particular setting of the subreddit where WritingPrompts was collected from; additionally, the average sentence lengths for the datasets SBERT were trained on were 14.1 and 22.3 [2, 18] whereas the average length of sentences in the WritingPrompts dataset is 28.4 [7]. We posit that SBERT has less exposure to longer, more complex (or ill-formed) sentences such as those found in WritingPrompts.

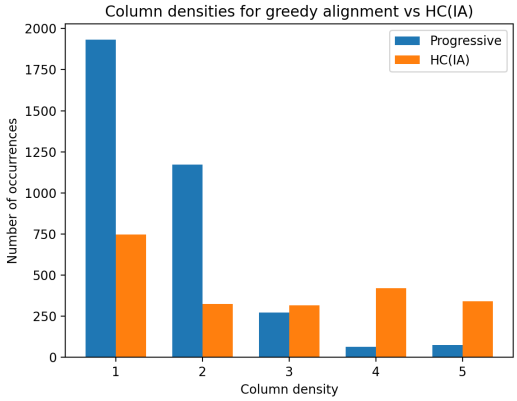

Figure 2: The progressive algorithm outputs alignments that are much sparser and longer than the best-performing method that optimizes the Steiner distance, hill climbing (IA). This is shown by a large number of columns from progressive alignments containing only a single non-gap element, and few columns containing more than 3 non-gap elements. These column density counts were obtained across all multiple alignments from the initial 50 clusters.

The result of this bias is that shorter sentences are more likely to be matched, and stories that contain many short sentences (especially those with many trivial nearest neighbours) may be favored.

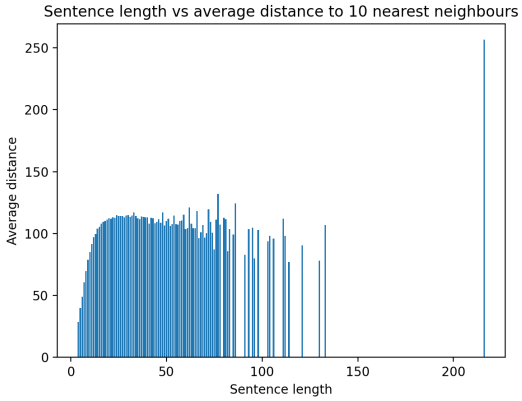

Figure 3: Longer sentences tend to be farther from their nearest neighbours. 2^{16} random sentences were sampled from WritingPrompts, and the plot shows their length versus the average distance to their 10 nearest neighbours. Distances were averaged across all sentences of the same length.

However, despite the sentence-level complexity of WritingPrompts, the stories themselves tend to have very simple structure. Due to their status as short stories, the traditional exposition, rising action, falling action, and conclusion segmentations do not fit very well. The short stories condense the structure and combine or completely leave out parts. Additionally, as we truncate stories at 25 sentences, we may miss the latter parts of the story, such as the conclusion and the falling action.

With the SBERT and data limitations in mind, we examine a particular set of 5 stories with nontrivial alignments as a case study. The prompt and first few sentences of each story are given in Figure 4. We find that the heuristics for finding similar sets of stories are effective. As the stories in our case study cluster are all about outer space, we refer to them as the space story cluster. The multiple alignment for the cluster, in Figure 5, although sparse, shows evidence of a joint storyline. Indices 1-14 of the alignment appear to be the introduction of stories 1, 2, 3, and 5, with story 4 exhibiting a less conventional structure and style. Due to their nature as short stories, much of the remainder of the stories fall under exposition.

We present a good sub-alignment in Figure 6, which involves indices 11 and 12 in Figure 5. This particular alignment occurs early in the multiple alignment, and captures shared structure across four of the five stories. These sentences mark the end of the introduction for each of their respective stories.

We also examine a poor sub-alignment in Figure 7, which involves indices 31-34 from the multiple

alignment, containing sentences from stories 4 and 5. The fragment from story 4 occurs towards the end that story's introduction, yet is aligned to sentences that occur after story 5's introduction. We hypothesize that story 4 is dissimilar to the other four stories in terms of structure and writing style, causing errors in the alignment.

The full multiple alignment for this cluster, as well as other clusters, can be viewed by following the instructions at the following site: github.com/dongruoping/story-clustering.

8 Discussion

We present a comparison of objectives and algorithms for finding multiple sequence alignments of stories, with the goal of inducing storylines. We find that progressive alignment results in the best alignments qualitatively. Analysis of the alignments obtained by the progressive algorithm discovered some structure in story clusters, supporting the hypothesis that storyline induction is possible without human annotation and with just a measure of semantic similarity.

Our goal is to develop a generative story model that explicitly models storylines by extracting them from multiple stories. The positive results of our experiments indicate progress towards this goal. For next steps, we will further explore and improve the three main components of our multiple sequence alignment pipeline: 1) the choice of story sets, 2) the similarity measure, and 3) the alignment algorithm itself.

Finding good story sets is integral to the method, as it is the first step in the approach, but is computationally expensive. Obtaining the best story set of a given size is similar to subset sum optimization, which is a difficult combinatorial optimization problem. Our approach relied on greedy heuristics to approximate the best story sets. In order to improve our approach, we will improve those heuristics while still relying on a greedy approximation. As a quick recap, our approach to story selection had two main steps: The first step was to find nearest neighbours using the bigram bag of sentence representations, and the second step was to rerank those nearest neighbours using pairwise alignment distances. The first step was necessary to obtain candidates for reranking, as computing the pairwise alignment distances for all pairs of stories would be too expensive. In order to improve the procedure, we will learn a measure that is better correlated with the pairwise alignment distance than the current measure involving bigram BoS representations. A simple method for accomplishing this would be to train a regression model on top of the bigram BoS representations to predict the distance obtained from pairwise alignment.

Improving the similarity measure requires designing a fine-tuning objective for story datasets in order to adapt the similiarity metric to the domain of interest. As SBERT, the basis of the similarity metric, was trained using human-labelled semantic matches for sentences, the same objective used to train SBERT cannot be used for story datasets without manual labeling. We instead propose a new generative model of multiple stories for fine-tuning, with the following generative process: First a multiple alignment is generated, a matrix whose entries are either a vector corresponding to a sentence embedding or a gap, then generate all sentences for each story conditionally independently from other sentences given the multiple alignment. This places the modeling burden on the prior distribution over multiple alignments. A first approach to modeling the multiple alignment would be to model it autoregressively from left to right, while the likelihood of a sentence given a sentence vector can be modeled with a sentence-level distribution [12]. Inference would entail inferring the multiple alignment given the stories, which we could approximate with a point estimate from the MSA procedure used in this work, allowing us to train the model via variational bayes EM.

For improving the alignment algorithm, we will explore continuous relaxations for multiple sequence alignment. Alayrac et al. [1] found that their method, which relied on a continuous relaxation and rounding procedure, outperformed off-the-shelf biological multiple sequence alignment libraries. Additionally, a method that admits a continuous relaxation may be more easily integrated within a neural network [15], allowing us to incorporate this procedure in improving the similarity measure as well.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Learning from narrated instruction videos. *CoRR*, abs/1506.09215, 2015. URL `http://arxiv.org/abs/1506.09215`.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326, 2015. URL `http://arxiv.org/abs/1508.05326`.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Khyathi Raghavi Chandu, Ruo-Ping Dong, and Alan W. Black. Reading between the lines: Exploring infilling in visual narratives. *CoRR*, abs/2010.13944, 2020. URL `https://arxiv.org/abs/2010.13944`.

[5] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *CoRR*, abs/1912.02164, 2019. URL `http://arxiv.org/abs/1912.02164`.

[6] Patrik D'haeseleer. What are dna sequence motifs? *Nature Biotechnology*, 2006. URL `https://rdcu.be/b9Co9`.

[7] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. *CoRR*, abs/1805.04833, 2018. URL `http://arxiv.org/abs/1805.04833`.

[8] Angela Fan, Mike Lewis, and Yann N. Dauphin. Strategies for structuring story generation. *CoRR*, abs/1902.01109, 2019. URL `http://arxiv.org/abs/1902.01109`.

[9] DF Feng and RF Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. 1987.

[10] Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2405. URL `https://www.aclweb.org/anthology/W19-2405`.

[11] Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2405. URL `https://www.aclweb.org/anthology/W19-2405`.

[12] Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. Toward better storylines with sentence-level language models, 2020.

[13] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs, 02 2020.

[14] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.174. URL `https://www.aclweb.org/anthology/2020.acl-main.174`.

[15] Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. Gradient estimation with stochastic softmax tricks, 2020.

[16] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011.

[17] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

[18] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426, 2017. URL http://arxiv.org/abs/1704.05426.

[19] Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. *CoRR*, abs/1811.05701, 2018. URL http://arxiv.org/abs/1811.05701.

Story 1
After hundreds of years of sending messages into the sky , humanity receives its first message from intelligent life . Decoded it simply says , " Be quiet before they find you . " " Zebin exclaimed as he received yet one more channel of communication from the Earth . Twenty years ago , the ambivalence over whether KIC 8462852 was in actuality an " alien mega structure " had finally come to an end after nearly 200 years of joint scientific endeavour by the leading lieges of the Earth . Since then , humanity had been trying with fervor to try and communicate with the star classified as a Dyson Sphere around 1480 light years away hoping that the far advanced civilisation might be generous enough to show the earthlings a way to solve their own energy crisis .
Story 2
Sunday Free Write : Leave A Story , Leave A Comment - New CSS Edition ! In 2056 NASA intercepted a frequency that was not of Earth . With its point of origin unknown they began to study it in an attempt to discover from whence it came . As it was studied it became known as the whoa signal , mockingly after the famous " wow !
Story 3
By 2345 humanity has colonized most of the solar system and have started sending probes beyond sol , when one day , a frantic interstellar message is received saying " do not leave the sol system " Listen well childen , for I have a Story of Old Earth to tell . Long ago , in ages past , the Men of ages past had but one True world , and had not yet learned to truely swim across the stars . In the nuturing embrace of Sol , humanity sowed life in the wake of the places Man gone .
Story 4
As the universe resonates for its final time , heat death approaching , the last energy of the universe is used to send a message . We knew something was going on , we knew something at any time would happen as time went on . Millennia after millennia we saw the lights fade out year by year pondering if new ones will be born to re-convey balance in the universe , but no the night sky kept getting darker star by star . We did n't care at first but the scientists all around our universe that we spread around knew something was wrong but were too busy with inorganically creating life on a possibly inhabitable planet around Zylon-B .
Story 5
Astronauts discover an abandoned space station orbiting a small moon . It is the far future . " Alright , we ' re here , " Captain Schiff said as the blocky craft approached the small disc-like object orbiting around Io , one of the five Galilean moons of Jupiter . An astronomer working for Global Colony Services , GSC , a young but highly profitable organization with fledgling colonies on the Moon , Mars , and in the asteroid belt , had spotted an irregular object in lunar-synchronous orbit that wasn ' t there last year . The GSC , or rather the league of nations sponsoring it , was concerned about a potential threat to the admittedly massive investments it had made in the asteroid belt , so it had sent the great ship MecaBubo 1 out to investigate .

Figure 4: The first few sentences from each of the 5 stories in our space story cluster analysis. The prompt is for each story is given in italics. All stories have a space-related theme. The first two sentences of story 5 have been filtered out, as they repeated the prompt.

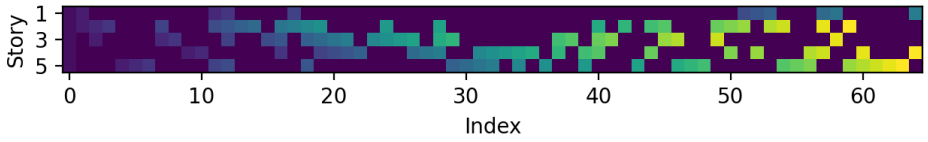

Figure 5: The multiple story alignment of the space story cluster, obtained via progressive alignment. We refer to the story in the top-most row as story 1, down through the bottom-most row as story 5. Each colored square corresponds to a sentence from the corresponding story, while a dark square corresponds to a gap. Sentences within a row proceed in order of their appearance in the story; only gaps are inserted within each row. An identical dummy beginning-of-story sentence token was inserted at the start of each story, yielding the full column of matches at index 0.

Story 1	**Story 2**	**Story 3**	**Story 5**
Twenty years ago , the ambivalence over whether KIC 8462852 was in actuality an " alien mega structure " had finally come to an end after nearly 200 years of joint scientific endeavour by the leading lieges of the Earth .	NASA discovered the signal was encrypted like nothing they had ever dreamed of ; the discovery of the encryption itself set technology hundreds of years ahead of where it once was .		An astronomer working for Global Colony Services , GSC , a young but highly profitable organization with fledgling colonies on the Moon , Mars , and in the asteroid belt , had spotted an irregular object in lunar-synchronous orbit that wasn ' t there last year .
Since then , humanity had been trying with fervor to try and communicate with the star classified as a Dyson Sphere around 1480 light years away hoping that the far advanced civilisation might be generous enough to show the earthlings a way to solve their own energy crisis .	It sparked the golden age of exploration in our solar system ; Ceres , Vesta , Hektor , Thisbe , Diotina , Fortuna were among many asteroids in the asteroid belt that were to be mined and inhabited ; the once failed colonization of Mars was reattempted and achieved , Europa of Jupiter , Titan of Saturn and Triton of Neptune all were to be colonized and inhabited ; Man had even reached as far as the Oort Cloud in the outer reaches of our solar system as early as 2096 .	From the Illiuminious research Stations of Mercury , to the Ruins of Old Chicago , Humanity had taken to the space around Sol with great bravado .	The GSC , or rather the league of nations sponsoring it , was concerned about a potential threat to the admittedly massive investments it had made in the asteroid belt , so it had sent the great ship MecaBubo 1 out to investigate .

Figure 6: An example of a good alignment between 4 story segments from the multiple alignment. The first sentences from stories 1, 2, and 5 detail an event which sparked the following sentences. The next sentences from all four stories mention a form of exploration, investigation, or expansion.

Story 4	**Story 5**
We tried making stars but they all died out faster then the stars above but it seemed somehow to make our time last just a little bit longer .	Not even great technicians at that, since the ship could handle damn near everything itself .
We had time to formulate whats going on and a joint meeting of scientists from all around the corners of the universe .	We were there as human eyes and ears , witnesses to send back any special details , in case they were necessary , and mechanics to fix the ship in case one of those " one in a million " scenarios cropped up (which seemed to be happening more and more often since children of Adam and Eve began to colonize their small corner of the cosmos .
Some full robotic now , half robotic and humanoid , one was full humanoid from Earth 1B and some that looked so foreign from living on a planet with no sunlight for 3 years at a time with the vegetation growing above 40 feet in the air .	
Enough about the scientists though .	Schiff was our leader , and a good enough man , if a bit formal .

Figure 7: An example of a poor alignment between two story segments from the multiple alignment. The sentences are weak semantic matches, although both segments serve as exposition in their respective stories.