Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2022)

Salience Vectors for Measuring Distance between Stories

Rachelyn Farrell,[1] Mira Fisher, [2] Stephen G. Ware [3]

Dept. of Computer Science, University of Kentucky, Lexington, KY USA
[1]rachelyn.farrell@uky.edu, [2]ocfi222@uky.edu, [3]sgware@cs.uky.edu

Abstract

Narrative planners generate sequences of actions that represent story plots given a story domain model. This is a useful way to create branching stories for interactive narrative systems that maintain logical consistency across multiple storylines with different content. There is a need for story comparison techniques that can enable systems like experience managers and domain authoring tools to reason about similarities and differences between multiple stories or branches. We present an algorithm for summarizing narrative plans as numeric vectors based on a cognitive model of human story perception. The vectors encode important story information and can be compared using standard distance functions to quantify the overall semantic difference between two stories. We show that this distance metric is highly accurate based on human annotations of story similarity, and compare it to several alternative approaches. We also explore variations of our method in an attempt to broaden its applicability to other types of story systems.

1 Introduction

Plan-based models of narrative are popular in games and other interactive systems for tracking causal connections between events and ensuring plot consistency. Few methods have been proposed for summarization and comparison of narrative plans, which could be useful in a variety of ways. Experience managers could make more informed decisions by considering differences and similarities between trajectories of an ongoing story (Jones and Isbell 2014; Amos-Binks, Potts, and Young 2017). Story distance metrics that are based on human perception may improve the accuracy of measurements such as plan-set diversity, which is often used to evaluate narrative planning models and algorithms (Porteous, Charles, and Cavazza 2013; Farrell and Ware 2016; Porteous et al. 2020). Our own focus is toward authoring tools, which could provide better feedback about the story model being created by clustering stories into similar groups. For this to be effective, we need a distance metric that considers similarities between the stories that are meaningful to humans.

We propose a story distance measurement using a novel story summarization technique based on a previously validated cognitive model. Our method encodes information about stories as numeric vectors that can be compared using standard distance calculations for an accurate measurement of the stories' overall difference. We test the accuracy of our distance metric and several alternative approaches in a human subjects evaluation. We find that our metric is at least as accurate as all the other methods we tried, and significantly more accurate than the other summary-based method. Although it is designed for a specific planning model, we believe that our approach is flexible and can be adapted to other kinds of story systems. We attempt to justify this by testing several variations of our model's definitions and showing that they all achieve similarly high accuracy. Finally, we release our dataset of human annotations of story similarity to facilitate future research in this area.

2 Related Work

Our technique builds on prior research in adapting the Event-Indexing Situation Model (Zwaan and Radvansky 1998) for a variety of uses in narrative planning systems. Indexter (Cardona-Rivera et al. 2012) measured the salience of past events based on their relatedness to the present event, where events can be linked through five indices: protagonist, time, space, intentionality, and causality. Later studies found that choices are perceived as more meaningful when their implied outcomes do *not* share any of these indices (Cardona-Rivera et al. 2014); and that readers preferred story endings that shared indices with prior choice outcomes (Farrell, Ware, and Baker 2020). These studies demonstrate that by tracking how events are linked in memory, we can model various aspects of the reader's perception of the story. In the present work we use similar constructs but for a unique goal: to model the reader's memory of complete or partial stories. The idea is that by modeling what people remember about each story and comparing these memories, we are approximating the process humans undertake when they mentally make the same comparison.

There is a large body of work on plan-based narrative models, which adapt classical planning structures and algorithms for specific purposes related to modeling stories (Young et al. 2013). This includes models of character intentionality that govern how characters may act in stories based on their goals (Riedl and Young 2010; Teutenberg and Porteous 2013; Ware and Young 2014). The only work we are aware of that seeks to summarize the narrative plans produced by these frameworks is the Important Step, Intention Frame (ISIF) summary (Amos-Binks, Roberts, and Young 2016). ISIF summarizes a plan as two sets: its important steps (having most causal connections to other steps) and its intention frames (the plans representing how characters intend to achieve their goals). For a distance measurement, the authors propose a Jaccard-based comparison of the overlap between these two sets. The accuracy of the ISIF distance metric according to humans has not been previously evaluated. Our approach differs from ISIF in that we are modeling more narrative features, and are using a vector representation that captures more nuanced information.

Common approaches to measuring the difference between a pair of classical plans include the set difference between the actions of the plans (Srivastava et al. 2007), and the edit distance, or number of editing operations required to convert one plan into the other (Levenshtein 1966). These methods can be applied to story plans, but since they only capture syntactic differences, they may not always agree with human assessments of story similarity. There are also linguistic approaches for comparing story texts, e.g. BLEU and ROUGE (Papineni et al. 2002; Lin 2004); these can be applied to story plans that have been translated into English sentences. Additionally, language transformers like BERT (Devlin et al. 2019) project a story text into a latent space where distance can be measured. We include several of these techniques in our evaluation to compare the accuracy of our distance metric to alternative methods. These are methods for measuring distance, but they do not summarize the content of stories. Our method is unique; it assigns a numeric value to the salience of each important story element, which can serve to summarize the story's content.

3 Methodology

In this work we represent stories as narrative plans using a STRIPS-like planning domain (Fikes and Nilsson 1972), where steps in the plan are grounded instances of parameterized operators, or actions. The following examples are from the *Grammalot* domain, which we adapted from a previous interactive narrative experiment (Ware et al. 2019). It contains seven actions: *walk*, *buy*, *attack*, *rob*, *loot*, *arrest*, and *wait*. It concerns the character Tom and his quest to get a potion and bring it back home without being attacked by a bandit or arrested by a guard in the process. The full domain is described in Appendix A.

Table 1 shows two *Grammalot* solutions, labeled X and Y. In X, Tom walks to the crossroads in the daytime and gets attacked by the bandit. In Y, Tom waits for night, then walks to the crossroads at nighttime and gets attacked by the bandit. Each event is represented as a grounded action containing typed parameters. In the $walk$ operator, the first parameter is a $character$, the second and third are $locations$, and the fourth is a $timeframe$. The arguments used to instantiate a specific instance of the operator are constants, e.g. Tom, $Cottage$, $Crossroads$, and Day for the action labeled a_{1X}.

Story	Action Signature	Label
X	walk(Tom, Cottage, Crossroads, Day)	a_{1X}
	attack(Bandit, Tom, Crossroads, Day)	a_{2X}
Y	wait(Tom, Cottage, Day)	a_{1Y}
	walk(Tom, Cottage, Crossroads, Night)	a_{2Y}
	attack(Bandit, Tom, Crossroads, Night)	a_{3Y}

Table 1: Example stories X and Y

Operators include preconditions, which must be true in the story world for the action to be applicable; and effects, which become true when the action is applied. The action a_{1X} changes Tom's location from the cottage to the crossroads: its precondition specifies $location(Tom) = Cottage$, and its effect sets $location(Tom) = Crossroads$.

We assume that the planning domain uses at least the three types mentioned: $character$, $location$, and $timeframe$, and that each operator contains at least one $location$ parameter and at least one $timeframe$ parameter. We explain the purpose of these assumptions in the definitions below. As we discuss in a later section, our model may still be usable with other story representations. As long as the entities below can be defined in some way, we believe this planning representation is not essential.

3.1 Salience Vectors

We summarize a story as a set of five fixed-length numeric vectors, which we call the *salience vectors*, because they represent the relative salience of different entities in the story at a given time. Each vector reflects one of the five situational dimensions defined in (Cardona-Rivera et al. 2012). Prior work deals with salience of *events* based on links through these dimensions, but here we model the salience of the *entities* involved in those links. For example, if two events are linked through the protagonist index, meaning they have a character in common, then instead of modeling the earlier *event* becoming salient when the later event occurs, we represent the *character* they have in common becoming salient itself. Below we describe the five dimensions, identify their associated entity types, and define which entities of each type become salient based on an action occurrence.

Protagonist The *protagonist* dimension links events that involve the same important characters. Since intentional planners explicitly distinguish between characters who are responsible for taking the action (the *actors*, for whom the planner must justify the action) and those who are involved in it but not willfully (e.g. a recipient of the action), we use this information in our definition.

Definition 1. When an action occurs, all characters who are actors (intentional participants) in the action become salient.

Time The *time* dimension links events that occur within the same time frame. We assume the planning domain explicitly provides the time frame of each operator in its parameters.

Definition 2. When an action occurs, all time frames among its arguments become salient.

Space The *space* dimension links events that occur in the same location. As with time frames, we assume locations are provided as arguments of each action.

Definition 3. When an action occurs, all locations among its arguments become salient.

Intentionality The *intentionality* dimension links events that occur for the same reason. Intentional planners differ in their representations of character goals and how they decide which goal an action is contributing to—e.g. intention frames (Riedl and Young 2010), relevant actions (Teutenberg and Porteous 2013), explanations (Ware and Young 2014). We use a model that justifies character actions using *explanations*: an explanation is a plan for a particular goal that a character can foresee at a given time (Shirvani, Farrell, and Ware 2018). We use the goals of an action's explanations to indicate the reasons why the action occurred.

This is not perfect: Some planners may not distinguish between *which* of a character's goals an action is in service toward, only that it is justified. Even when this information is modeled by the planner, it is still possible that there are multiple valid explanations for the character's action, or that the audience could interpret the behavior in a completely different way. We use this limited definition for now but hope that future work will improve upon it, e.g. through goal recognition techniques.

Additionally, we extend this definition to include *motivated* goals. We say that an action motivates a goal when it changes the goal condition from being satisfied (true) to unsatisfied (false). For example, in the *Grammalot* domain there is a guard who wants to punish criminals: He has a goal for each other character that they are either not a criminal, or have been punished. Initially this is true for Tom: He is not a criminal. If Tom commits a crime, it becomes false, so this motivates the guard's goal to punish Tom.

Definition 4. When an action occurs, all character goals used in explanations for that action, or that are motivated by the action, become salient.

Causality The *causality* dimension links events that are causally related, meaning the earlier action enables the later action. Planners can readily access this information through causal links or equivalent structures: when an action a has an effect p, a later action a' has a precondition p, and no action in between a and a' negates p, there is a *causal link* from a (the parent) to a' (the child). An action's causal ancestors are its causal parents and all actions in the transitive closure of this relation.

Definition 5. When an action occurs, the action itself and all its causal ancestors become salient.

Algorithm Algorithm 1 summarizes a given story with a set of five vectors representing the salience of all entities at the end of the story. For our work we are interested in comparing complete stories, so we capture summaries at the end. To summarize incomplete stories we would simply halt this algorithm at the desired step. The function SALIENT($entity$, $action$) returns whether an entity is made salient by a particular action, based on its type, according to the definitions given above.

Algorithm 1: Create salience vectors for a given story

```
procedure SUMMARIZE($story$, $d$)
    $E \leftarrow \{C, T, L, G, A\}$        ▷ the set of sets of entities
    $V \leftarrow \{v_1, v_2, v_3, v_4, v_5\}$        ▷ such that $|v_i| = |E_i|$
    Initialize all values in $v_{1...5}$ to 0.
    for $action$ in $story$ do
        for $i$ in 1...5 do
            for $j$ in 1...$|E_i|$ do
                if SALIENT($E_{ij}, action$) then
                    $V_{ij} \leftarrow 1$
                else
                    $V_{ij} \leftarrow V_{ij} * d$
```

Let d be a constant decay factor between 0 and 1 (we use 0.5 as a default). Let C be the set of characters in the domain, T the set of time frames, L the set of locations, G the set of character goals, and A the set of grounded actions. We define the set of entities E comprising these five sets. The algorithm begins by initializing the set of salience vectors V with five numeric vectors, each having length equal to the number of elements in the corresponding set in E.

Entities are assigned a zero salience value initially, representing no salience at all. The procedure then steps through the actions in the story, assigning the maximum salience value (1) to each entity involved in the current action, and decreasing the salience of all entities that are not involved. The resulting set of vectors V represents the salience of each entity in E at the end of the story.

Table 2 shows the salience vectors after each step in story X from Table 1, using $d = 0.5$. Tom's goal and the bandit's goal are abbreviated as g_T and g_B, respectively, and actions using the labels shown in Table 1. Some entities whose values remain zero across all vectors are omitted for space.

After step 1, the entities *Tom*, *Day*, *Cottage*, *Crossroads*, Tom's goal (to get home with the potion), and the action a_{1X} itself—*walk(Tom, Cottage, Crossroads, Day)*—are salient. After step 2, *Bandit*, the bandit's goal, and the new action a_{2X} have become salient. Entities that are still involved remain salient: *Day*, *Crossroads*, and a_{1X} (since it is a causal ancestor of a_{2X}). Uninvolved entities have decayed: *Tom* (since he was not an actor in the attack), *Cottage*, and Tom's goal. The final row reveals a great deal of information about the story; e.g. that it ends with the bandit acting on his goal at the crossroads, that it never involves night or the other locations and characters (not shown), and which specific actions are important.

Table 3 shows the vectors for the second story (Y); we can see that it, too, ends with the bandit acting on his goal at the crossroads, but at night and with different actions. The vectors highlight some similarities between the two stories as well as some differences; they are similar in terms of their characters, locations, and goals (v_1,v_3, and v_4), but different in their time frames and causal structure (v_2 and v_5).

The fact that the salience vectors contain information

Step	v_1 (characters)		v_2 (times)		v_3 (locations)		v_4 (goals)		v_5 (actions)				
	Tom	$Bandit$	Day	$Night$	$Cottage$	$Crossroads$	g_T	g_B	a_{1X}	a_{2X}	a_{1Y}	a_{2Y}	a_{3Y}
1	1	0	1	0	1	1	1	0	1	0	0	0	0
2	0.5	1	1	0	0.5	1	0.5	1	1	1	0	0	0

Table 2: Salience vectors after each step in story X

Step	v_1 (characters)		v_2 (times)		v_3 (locations)		v_4 (goals)		v_5 (actions)				
	Tom	$Bandit$	Day	$Night$	$Cottage$	$Crossroads$	g_T	g_B	a_{1X}	a_{2X}	a_{1Y}	a_{2Y}	a_{3Y}
1	1	0	1	0	1	0	1	0	0	0	1	0	0
2	1	0	0.5	1	1	1	1	0	0	0	1	1	0
3	0.5	1	0.25	1	0.5	1	0.5	1	0	0	1	1	1

Table 3: Salience vectors after each step in story Y

about the content of the stories is a key benefit of our methodology, and an important advantage of our distance metric over the others we evaluate. The salience of story entities can be used in a variety of ways beyond the calculation of a distance measurement. For example, it can be used to convey *what* is different between stories, or to describe a particular group of stories in terms of what makes them similar or sets them apart from the rest, as well as to summarize information about solution spaces as a whole. Specific techniques for accomplishing these goals are not addressed in this paper; instead we focus on evaluating the accuracy of our distance metric compared to others.

3.2 Distance Calculation

We define the salience distance SD between two stories X and Y as a linear combination of the distances between their corresponding salience vectors of each type:

$$SD(X, Y) = \sum_{i=1}^{5} w_i * NSE(v_{iX}, v_{iY}) \qquad (1)$$

where w_i are relative weights for each dimension, which sum to 1 (by default we use equal weights, $w_* = 0.2$), and $NSE(u, v)$ calculates the normalized squared Euclidean distance between a pair of vectors (their squared Euclidean distance after scaling their lengths to have unit norm). This is an appropriate function for this case because the magnitudes of the vectors are not important; only their directions. This accounts for the unbalanced sizes of the different vectors (e.g. the causality vector is likely to be much larger than the other four). The normalized squared Euclidean distance is always between 0 and 1, so the final salience distance value is also bound between 0 and 1.

We calculate the salience distance between stories X and Y from Table 1 using Equation 1, with $v_{1\ldots5X}$ from the bottom row of Table 2, and $v_{1\ldots5Y}$ from that of Table 3. Since the final v_1, v_3, and v_4 happen to be identical (characters, locations, and goals are equally salient at the end of both stories), these contributions to the equation are zero. The differences come only from the time and causality vectors:

$$\begin{aligned} SD(X, Y) =& 0.2 * NSE([1, 0], [0.25, 1]) \\ &+ 0.2 * NSE([1, 1, 0, 0, 0, \ldots], [0, 0, 1, 1, 1, \ldots]) \end{aligned}$$

The result is a distance value of 0.296, which is relatively small, as we imagine it should be. Tom either waits for night or does not, but otherwise the stories are the same. Notice that these two stories share no identical action signatures, so metrics like ISIF would consider them highly different. Our metric is more nuanced; it captures the similarity of some elements while recognizing differences in others.

3.3 Variations

To demonstrate the model's flexibility we test several variations of the definitions given in Section 3.1. Our protagonist definition distinguishes between actors and non-actors, but this is not strictly necessary. We test two alternative definitions: In one, *all characters* in the action's arguments become salient, regardless of their intent. In the other, only the predefined *story protagonist*—Tom in this case—becomes salient when present in the action's arguments. We also test an alternative version of intentionality in which motivated goals are not considered; only the goals used in explanations for the action become salient.

Finally, we test two alternative representations of causality. Properties in our planning system are represented as assignments in the form $fluent = value$. For example, Tom's location is a fluent, $location(Tom)$, and the assignment $location(Tom) = Cottage$ states that Tom is at the Cottage. One alternative definition for causality uses the set of all unique *fluents* as entities, rather than grounded actions; and the other, the set of all unique *assignments*. In both cases, a fluent or assignment becomes salient if it is used in the causal ancestry of the current action, meaning it is the property that establishes a causal link to the action or one of its causal ancestors. We did not vary time or space due to the lack of suitable alternative definitions.

We tested all 18 possible combinations of definitions. The results in Section 5 show only the best and worst scores among these. The best variation was a tie between the two that used our original protagonist definition with fluents for causal entities (with and without motivated goals). The worst variation used the story protagonist definition with assignments for causality, and included motivated goals. Results for all variations are included in Appendix B.

3.4 Parameters

We tested all possible combinations of the weights $w_{1...5}$ ranging from 0 to 1 in increments of 0.1, using the default decay rate $d = 0.5$ and the default vector definitions (Section 3.1). The results in Section 5 show the best and worst scores achieved using modified weights. The weights that achieved the highest scores are shown in Appendix C. They were different between the two analyses and there were many ties, but one clear trend is that time should be weighted low or zero for the most accurate performance in this domain. We also tested different decay rates ($0.05 \leq d \leq 0.95$, in increments of 0.05) using the default weights and vector definitions. Scores for the best and worst decay are also shown in Section 5 (best ≤ 0.25, worst $= 0.95$).

4 Evaluation

Since we are comparing metrics with different scales, we cannot directly compare their distance values. Instead we use *comparisons* between two values: A comparison is a triplet of solutions $\langle REF, A, B\rangle$ where REF is a reference story and A and B are two other stories, and is interpreted as the question, "Which story is more similar to the reference: A, or B?" This condition can be evaluated by any distance matrix: If the distance between REF and A is less than the distance between REF and B, then the answer is A. If the two distances are equal, the answer is undefined; otherwise it is B. We use human answers as ground truth to obtain accuracy scores for each metric.

We completed a breadth-first search of the *Grammalot* planning problem up to depth 6, which produced 58 solutions ranging from 2 to 6 steps long. Subjects were recruited using the online crowd-sourcing platform Prolific. They read a description of the domain (Appendix A), then completed 8 tasks in which they read a reference story followed by a pair of stories (shown in random order). Subjects were randomly assigned a question type: whether we ask which story is "more similar to" or "more different from" the reference (and invert their answers accordingly). To limit cognitive load, the reference story remained the same for a given participant throughout the whole task, as did the question type. Equality was not an option; A and B were the only available answers.

The specific stories we used were randomly selected, constrained only on the requirement that each subject would only view one reference story. We first randomly sampled the solutions for 8 different stories to use as references, then randomly sampled different stories for A and B. We asked 64 subjects 8 questions each and collected 512 total answers.

We compare salience distance to the following metrics from the literature. We used each method to produce a distance matrix containing pairwise distance values for all 58 solutions. Several methods require text input; for these we first translated the solutions into basic English sentences (the same translations we showed to participants). The solutions and translations are included in the linked dataset.

Action Distance is the Jaccard distance between the sets of grounded actions in each plan (Srivastava et al. 2007). The Jaccard distance between two sets X and Y is:

$$J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

ISIF measures a combined Jaccard distance between two sets of grounded actions taken from each plan. The *important steps* (IS) of a story are those with the most causal connections to other steps. The *intention frame summaries* (IF) of a story are the motivating and satisfying steps of the plans characters enact in pursuit of their goals. The ISIF distance between two stories X and Y is defined as:

$$ISIF(X,Y) = 1 - \frac{1}{2}\Big(\frac{|IS_X \cap IS_Y|}{|IS_X \cup IS_Y|} + \frac{|IF_X \cap IF_Y|}{|IF_X \cup IF_Y|}\Big)$$

Edit Distance counts the number of insertions, deletions, or substitutions required to convert one plan into the other (Levenshtein 1966). We use three variations: *Edit Action* applies these operations to whole action signatures; *Edit Symbol* applies them at the symbol level (the action name and parameters); and *Edit Word* applies them at the word level using the English translation of the story set.

BLEU is an automatic evaluation metric for summary comparison that evaluates textual similarity as co-occurrence of subsequences between a target text and a reference text (Papineni et al. 2002).[1] BLEU considers a combined score from multiple lengths of n-grams; we tested all possible equally-weighted groupings of these n-gram lengths. To present the range, we report the lowest-scoring BLEU method (using only unigrams or unigrams and bigrams) and the highest score (e.g. using only trigrams—most combinations tied for this score).

Rouge is another text comparison metric similar to BLEU, but considers a single n-gram length or the longest common subsequence (Lin 2004). We tested Rouge using n-grams 1 through 5, and Rouge-L (longest common subsequence). The lowest Rouge score was a tie between Rouge-1 and Rouge-L; the highest between Rouge-3, Rouge-4, and Rouge-5.

BERT (Devlin et al. 2019) is an English language text transformer which encodes similarities of input texts by placing them in a latent space. We process each story as a sequence of sentences and extract the final predicted position of each sequence in latent space, then calculate the cosine distance between two of these positions for a distance measurement. We tested this with both an unmodified instance of the base transformer and five tuned instances. For these we processed the text stories into two datasets for the Next Sentence Prediction and Masked Language Modeling tasks, and trained BERT for these tasks for a number of training epochs (1...5).

[1] We use REF as the reference to which A and B are compared.

Note that this does not fine-tune BERT on the task of measuring distance between stories, but on generating text more like our stories. We report the lowest score (using the unmodified BERT), and the highest score (the instance with one training epoch).

5 Results

We compare the metrics in two ways; first using only the questions where people agreed on an answer, and second using all the questions. Subjects significantly agreed if at least 7 out of 8 provided the same answer (binomial test, $p = .039$), which occurred for 37 of the 64 questions. Figure 1a shows the results of the analysis using these questions, in which a metric scores one point for each question it answers the same as the majority of humans. Undefined answers (equal distances) are counted as incorrect.

A binomial test indicates that a score of 24 or higher is statistically better than chance. Most metrics exceed this threshold. The default variation of Salience answers 34 out of 37 questions correctly (92% accuracy), which is higher than all other metrics except the best BERT (with which it is tied) and the best Rouge. To assess whether any of these differences are significant, we use a z-test of two proportions. The dotted vertical lines indicate the thresholds where the test becomes significant: A score of 28 or below is significantly worse than 34, and only 38 or above, which is impossible, would be significantly higher. Salience performed very well in this test; it was significantly more accurate than ISIF and Edit Action, and scored higher than almost all other metrics, but not significantly higher.

We also measure accuracy using all 512 answers collected, scoring one point for each individual human answer the metric agrees with. This effectively weighs the score for each question by how strongly people agreed on it. We also award half a point for undefined answers, since in this case people did not necessarily agree that one answer is correct. This means that providing a wrong answer is worse than providing no answer; and that when people were evenly split (this happened for 5 questions), all three possible answers (A, B, and undefined) are worth the same 4 points. The highest possible score is 420, representing agreement with the majority on every question.

Figure 1b shows the results of this analysis. The default Salience scores 388, and again we show the low and high z-test thresholds, 360 and 414. Note that no p-value correction is being applied, so we might reasonably consider comparisons on the thresholds to be insignificant. Here the default Salience outperforms all other metrics except the best Rouge, and significantly outperforms the unmodified BERT and Edit Word in addition to Edit Action and ISIF.

Dataset The story similarity data can be downloaded at:

http://cs.uky.edu/~sgware/projects/storysimilarity/

6 Discussion and Limitations

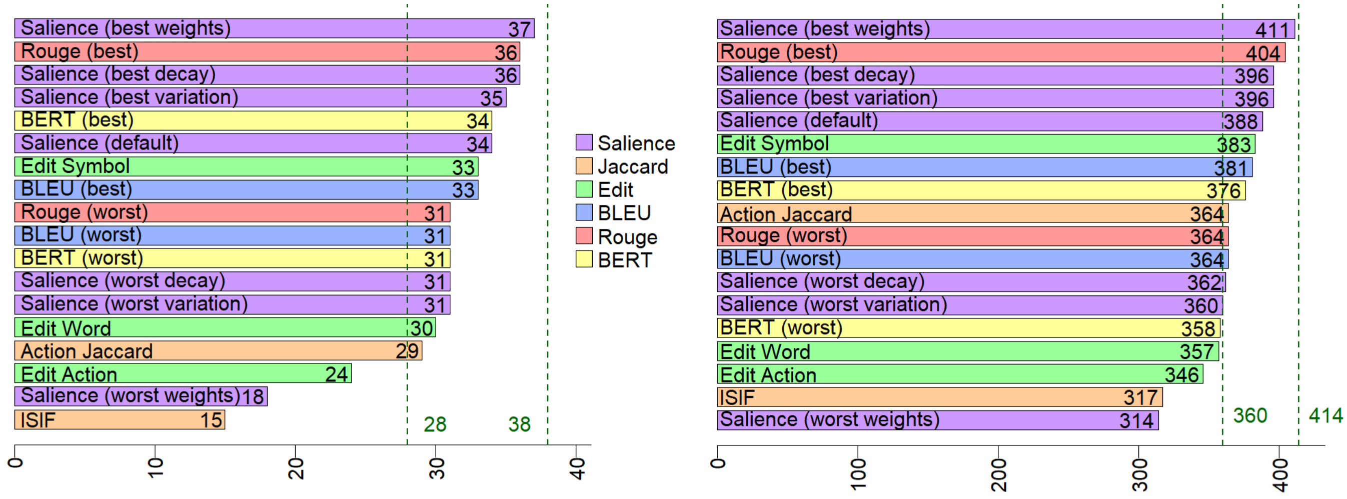
Both analyses suggest that intelligently choosing values for w_* and d may improve accuracy, although the improvements here were not significant. Equal weighting appears to be a good default, but better weights may be informed by domain knowledge (e.g. dimensions or entities that are more or less important). The ideal decay value is also likely to vary by domain, but may be related to measurable properties like average story length. None of the 18 variations of Salience were significantly better or worse than the default, with accuracy ranging from 84-95%. This suggests that our technique does not strictly depend on one set of definitions, and may be adaptable to other kinds of story systems.

While most metrics performed well in both analyses, ISIF is a notable exception. ISIF and Action Jaccard have a tendency to produce distance values of 1, e.g. for stories with no identical action signatures, like our example pair (Section 3). This results in many undefined answers, i.e. when both distance values are 1, which explains why they perform better in the second analysis than the first (undefined answers are worth half credit). Still it was surprising to see ISIF perform worse than Action Jaccard, upon which it is based. The difference is that Action Jaccard compares the full set of actions in both stories, whereas ISIF only compares actions of specific importance. These action sets may be helpful toward describing important story differences, but as a distance metric our study shows it to be less accurate than considering the full sets of actions.

The text-based methods generally performed well, with the best being Rouge-3+. This is a viable approach for measuring story distance, as long as the stories can be automatically translated into natural language. Note that these methods are sensitive to different text realizations; the translations we used are simple and repetitive, which is helpful to these metrics. Edit Symbol also performed well, and does not rely on text translations, but does require a plan-like symbolic representation. Salience is comparable in accuracy to these approaches, does not depend on text translations, and does not necessarily depend on a planning representation either (i.e. the relevant information could perhaps be extracted from text stories, or otherwise modeled differently from how we have defined it here). Furthermore, as noted previously, the primary benefit of salience distance is that the salience vectors capture meaningful content about the stories that can explain where their differences and similarities come from. The other methods, apart from ISIF, do not have this capability.

We believe this work may be useful for a variety of research in this field, but our specific purpose is to build authoring tools that can communicate the content of large solution spaces to domain authors. In the future we plan to cluster the set of stories and label the clusters using information from the salience vectors (e.g. the name of an entity that is always salient in this cluster and never in the others). This is why we prefer to use grounded actions as causal entities rather than fluents, even though fluents performed slightly better in our evaluation. We find that grounded actions make more intuitive cluster labels than fluents; they provide information in a form people expect to see. Fluents may be preferable for other purposes, however, since they capture more nuance than grounded actions.

We only used one domain for this study due to limited resources. This domain models enough variety to capture



(a) Using the 37 questions humans agreed on

(b) Using all questions

Figure 1: Accuracy Results

some differences between metrics, but perhaps we would see more significant differences in a larger domain. If more resources become available in the future, we would like to replicate these results in other domains and with more data, as this would allow us to draw stronger conclusions.

We showed that our distance metric is effective under varying definitions, but we did not do this exhaustively. These definitions are a starting point, but leave room for improvement, especially as planner intentionality models improve. Time and space could also be represented hierarchically. Furthermore it may be worth including other situational dimensions beyond the five we discuss, e.g. emotions, ideas, or objects.

What are we missing? Salience with best weights was the only metric to correctly answer all 37 questions on which subjects agreed, but this is not a fair metric since we cannot assume the optimal weights are known. The following is a question that no fair metric answered correctly, despite subjects agreeing on its answer with a 7/8 majority.[2]

> (Reference) Tom walks to crossroads. Bandit walks to market. Guard walks to crossroads. Bandit walks to crossroads. Bandit attacks Tom.
>
> (A) Tom walks to crossroads. Tom waits for night. Bandit attacks Tom.
>
> (B) Tom walks to crossroads. Tom walks to market. Tom buys potion. Bandit walks to market. Bandit attacks Tom.

According to the majority of our subjects, story A is more similar to the reference than story B is; yet almost all of the metrics answered that B is more similar. B contains more of the same actions and content as the reference compared to A, so why do humans think A is more similar?

In the reference, Tom goes to the crossroads and *does nothing,* while others walk in and out of the market, and then he gets attacked. It may seem like Tom bears some responsibility for this fate—if he had done something instead of nothing, he might have avoided it. Story A involves waiting for night, which is semantically similar to doing nothing. In A, Tom goes to the crossroads and *allows* himself to be attacked, while in B, he tries to achieve his goal but gets attacked in the process. In this light, B feels different, despite containing many of the same elements as the reference and being similar in structure and text.

Of course this is just one possible interpretation of why subjects chose A over B. It may be that waiting is similar to doing nothing, or perhaps this example reflects another type of similarity that we are not accounting for, e.g. thematic or high-level structural similarities. Whatever the case, this example demonstrates that some semantic similarities are not captured by any of these metrics. Future work may benefit from exploring ways to account for situations like this.

7 Conclusion

We have defined a numeric vector that summarizes a story's content by calculating the salience of each entity involved in the story upon its ending. Distances between these vectors can model the extent to which two stories are semantically different from each other. We showed that our distance metric is accurate according to humans in an example domain and compares favorably to existing methods for measuring story distance.

Our approach is unique in that the basis for the distance measurement is a summary of each story, based on a cognitive model of story comprehension, which contains information that can explain the stories' similarities and differences in detail. We find the results of this evaluation promising and intend to explore further applications of the salience vectors in future work.

[2] The story text has been abbreviated here for space.

A Domain Description

Tom needs a special potion from the merchant who works at the market. Tom's cottage is connected to the market via a crossroads which also connects to the merchant's house. There is a bandit lurking in the crossroads, who is a known criminal. There is a guard at the market who wants to punish criminals, but the guard does not initially know where the bandit is. At night, the merchant takes all her possessions to her house and goes to sleep.

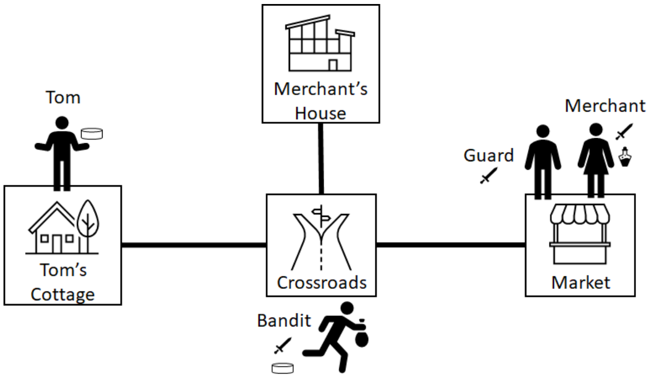


Figure 2: Initial state depiction

What characters want:

- Tom wants to have the potion and be back at his cottage.
- The merchant wants to sell her items for coins.
- The guard wants to punish criminals.
- The bandit wants to have valuable items (coins and potions).

What characters can do:

- Walk between connected locations
- Buy something from the merchant (for one coin)
- Attack someone (while armed with a sword)*
- Rob someone who is unarmed (while armed)*
- Rob someone who is asleep*
- Loot something from a dead person*
- Arrest a criminal (only the guard can do this)
- Wait for night (only Tom can do this)

* (Attacking and stealing are crimes.)

What characters have:

- Tom has one coin.
- The merchant has the potion and a sword.
- The guard has a sword.
- The bandit has a sword and one coin.

Stories end when Tom either dies, gets arrested, or returns to his cottage with the potion.[3]

B Accuracy for 18 Salience Distance Variations

The default variation is labeled *salience_default*. The other variations are named to indicate which alternative definitions are used: *c=F* means it uses fluents for causality; *c=A* means assignments for causality; *p=A* means it uses the "all characters" definition of protagonist; *p=S* means the story protagonist definition; and *!M* means it uses the version of intentionality that does not include motivations.

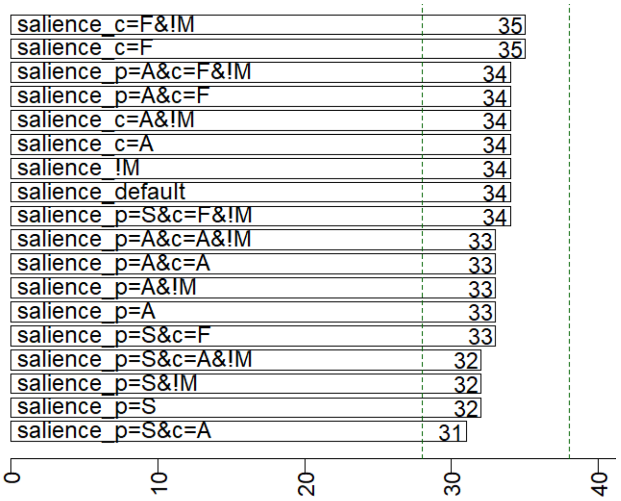


Figure 3: Using the 37 questions humans agreed on

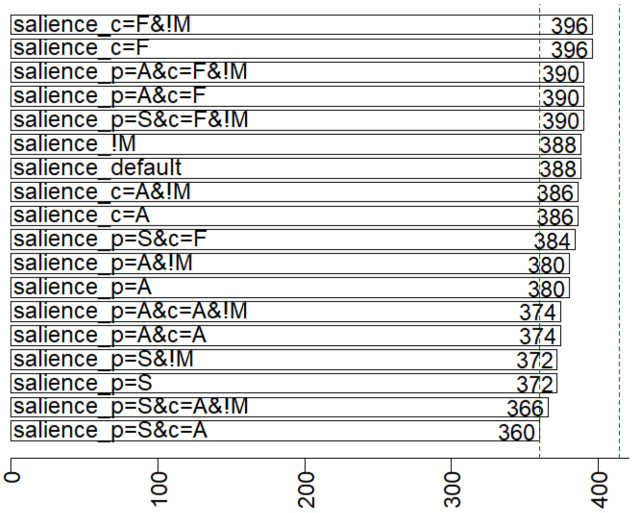


Figure 4: Using all questions

[3] These are the end conditions that constitute solutions to the planning problem.

C Best Performing Weights

Protagonist	Time	Space	Intentionality	Causality
0.0	0.0	0.1	0.3	0.6
0.0	0.0	0.1	0.4	0.5
0.0	0.0	0.1	0.5	0.4
0.0	0.0	0.1	0.6	0.3
0.0	0.0	0.2	0.3	0.5
0.0	0.0	0.2	0.4	0.4
0.0	0.1	0.2	0.4	0.3
0.0	0.1	0.2	0.5	0.2
0.0	0.1	0.3	0.4	0.2
0.0	0.1	0.3	0.5	0.1
0.0	0.1	0.3	0.6	0.0
0.0	0.1	0.4	0.5	0.0
0.1	0.0	0.1	0.2	0.6
0.1	0.0	0.1	0.3	0.5
0.1	0.0	0.1	0.4	0.4
0.1	0.0	0.1	0.5	0.3
0.1	0.0	0.2	0.2	0.5
0.1	0.0	0.2	0.3	0.4
0.1	0.1	0.2	0.3	0.3
0.1	0.1	0.2	0.4	0.2
0.1	0.1	0.2	0.5	0.1
0.1	0.1	0.2	0.6	0.0
0.1	0.1	0.3	0.4	0.1
0.1	0.1	0.3	0.5	0.0
0.2	0.0	0.1	0.2	0.5
0.2	0.0	0.1	0.3	0.4
0.2	0.0	0.1	0.4	0.3
0.2	0.0	0.2	0.2	0.4
0.2	0.1	0.2	0.3	0.2
0.2	0.1	0.2	0.4	0.1
0.2	0.1	0.2	0.5	0.0
0.2	0.1	0.3	0.4	0.0
0.3	0.0	0.1	0.2	0.4
0.3	0.0	0.1	0.3	0.3
0.3	0.1	0.2	0.4	0.0
0.3	0.1	0.3	0.3	0.0
0.4	0.0	0.1	0.2	0.3
0.4	0.1	0.2	0.3	0.0
0.4	0.1	0.3	0.2	0.0
0.5	0.1	0.2	0.2	0.0

Table 4: Weights scoring 37 on the first analysis

Protagonist	Time	Space	Intentionality	Causality
0.5	0.0	0.1	0.0	0.4
0.6	0.0	0.1	0.0	0.3

Table 5: Weights scoring 411 on the second analysis (both scored 36 on the first)

References

Amos-Binks, A.; Potts, C.; and Young, R. M. 2017. Planning graphs for efficient generation of desirable narrative trajectories. In *Proceedings of the 13th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, volume 13.

Amos-Binks, A.; Roberts, D. L.; and Young, R. M. 2016. Summarizing and comparing story plans. In *Proceedings of the 7th Workshop on Computational Models of Narrative*.

Cardona-Rivera, R. E.; Cassell, B. A.; Ware, S. G.; and Young, R. M. 2012. Indexter: a computational model of the Event-Indexing Situation Model for characterizing narratives. In *Proceedings of the 3rd Workshop on Computational Models of Narrative*, 34–43.

Cardona-Rivera, R. E.; Robertson, J.; Ware, S. G.; Harrison, B.; Roberts, D. L.; and Young, R. M. 2014. Foreseeing meaningful choices. In *Proceedings of the 10th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 9–15.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 4171–4186.

Farrell, R.; and Ware, S. G. 2016. Fast and diverse narrative planning through novelty pruning. In *Proceedings of the 12th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 37–43.

Farrell, R.; Ware, S. G.; and Baker, L. J. 2020. Manipulating narrative salience in interactive stories using Indexter's Pairwise Event Salience Hypothesis. *IEEE Transactions on Games*, 12(1): 74–85.

Fikes, R. E.; and Nilsson, N. J. 1972. STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3): 189–208.

Jones, J. K.; and Isbell, C. L. 2014. Story similarity measures for drama management with TTD-MDPs. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 77–84.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, 707–710.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81. Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318.

Porteous, J.; Charles, F.; and Cavazza, M. 2013. NetworkING: using character relationships for interactive narrative generation. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 595–602.

Porteous, J.; Ferreira, J. F.; Lindsay, A.; and Cavazza, M. 2020. Extending Narrative Planning Domains with Linguistic Resources. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*.

Riedl, M. O.; and Young, R. M. 2010. Narrative planning: balancing plot and character. *Journal of Artificial Intelligence Research*, 39(1): 217–268.

Shirvani, A.; Farrell, R.; and Ware, S. G. 2018. Combining intentionality and belief: revisiting believable character plans. In *Proceedings of the 14th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 222–228.

Srivastava, B.; Nguyen, T. A.; Gerevini, A.; Kambhampati, S.; Do, M. B.; and Serina, I. 2007. Domain Independent Approaches for Finding Diverse Plans. In *Proceedings of the 2007 International Joint Conference on Artificial Intelligence*, 2016–2022.

Teutenberg, J.; and Porteous, J. 2013. Efficient intent-based narrative generation using multiple planning agents. In *Proceedings of the 2013 international conference on Autonomous Agents and Multiagent Systems*, 603–610.

Ware, S. G.; Garcia, E. T.; Shirvani, A.; and Farrell, R. 2019. Multi-agent narrative experience management as story graph pruning. In *Proceedings of the 15th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 87–93.

Ware, S. G.; and Young, R. M. 2014. Glaive: a state-space narrative planner supporting intentionality and conflict. In *Proceedings of the 10th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 80–86.

Young, R. M.; Ware, S. G.; Cassell, B. A.; and Robertson, J. 2013. Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2): 41–64.

Zwaan, R. A.; and Radvansky, G. A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2): 162.