Calibrating a Metric for Similarity of Stories against Human Judgment *

Raquel Hervás, Antonio A. Sánchez-Ruiz, Pablo Gervás, Carlos León

Dep. Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid (Spain)
raquelhb@fdi.ucm.es, antsanch@fdi.ucm.es, pgervas@sip.ucm.es,
cleon@fdi.ucm.es

Abstract. The identification of similarity is crucial for reusing experience, where it provides the criterion for which elements to reuse in a given context, and for creativity, where generation of artifacts that are similar to those that already existed is not considered creative. Yet similarity is difficult to compute between complex artifacts such as stories. The present paper compares the judgment on similarity between stories explained by a human judge with a similarity metric for stories based on plan refinements. The need to identify the features that humans consider important when judging story similarity is paramount on the road to selecting appropriate metrics for the various tasks.

Keywords: similarity, novelty, stories, plans.

1 Introduction

Appropriate metrics for similarity are fundamental tools in many fields of Artificial Intelligence. For instance, there are several data mining and machine learning methods that are based on the similarity between the elements being considered. In case-based reasoning, similarity metrics are crucial for the retrieval and reuse of previous cases. Similarity is also fundamental for computational creativity because artifacts that are very similar to previously existing ones might not be considered creative. For this reason, it is important to take into account whether the metrics considered for a particular task adequately represent the concept of similarity that humans faced with the same task would apply. The present paper compares the judgment on similarity between stories explained by a human judge with a particular similarity metric for stories. The main goal is to identify which of the features that a human considers when evaluating story similarity are already taken into account by the metric, and which ones are not. The results of this comparison should provide a check list that might later on be applied to evaluate the appropriateness of other metrics.

* The research reported in this paper was partially supported by the Project WHIM 611560 funded by the European Commission, Framework Programme 7, the ICT theme, and the Future and Emerging Technologies FET programme; and by the Spanish Ministry of Economy and Competitiveness under grant TIN2014-55006-R.

We focus on the structural similarity of stories represented as plans composed of actions corresponding to the events in the story. In order to do so, we apply a similarity metric based on plan refinements and compare the obtained results for a pair of stories with the similarities found by a human expert. The key point of this comparison is that the metric does not only calculate a numerical similarity between the compared stories, but provides a report of the found similarities. This report is then compared with the observations obtained by the human expert. The comparison allows us to see if the automatic metric has been able to grasp the same features the expert considered important, and if structural similarity is enough for comparing computer-generated stories.

2 Previous Work on Similarity for Stories

Existing work on similarity for stories has focused on two different axes: story similarity for retrieval and classification of stories, and story similarity applied to the assessment of their novelty in a computational creativity setting.

2.1 Similarity Metrics for Story Generation

In general, there is relative consensus on the fact that comparing stories can be made at different levels. Comparing stories at a relatively abstract level is common, to the point of comparing not the exact sequence of events but the overall plot, or even the relations between the characters. This aspect of narrative has been addressed by structuralist and cognitive Narratology.

In particular, comparing narratives has been a long term goal of Computational Narrative, and several approaches have been taken with varying results [2, 10, 8]. Different aspects beyond pure literary composition have been tackled: structure alignment in bioinformatics [1], event mapping [3], and other approaches like considering story similarity in terms of the common summary that might be abstracted from the two stories being compared [9].

2.2 Similarity Metrics for Assessing Novelty of Stories

With respect to the assessment of creativity, a fundamental pillar is whether the results of a creative process have produced novel artifacts [14]. Research on the evaluation of creativity has addressed this point as an important requirement for the scientific exploration of creativity, and an important one for computational approaches. In [11], novelty of a given story is assessed in terms of new elements that appear in the story, or instances where existing elements have been replaced by elements of a different type. In [12], novelty of stories is considered in terms of their differences with an initial set of reference stories, based on the sequence of actions, the structure of the story in terms of emotional relations and tensions between the characters, and the occurrence of repetitive patterns.

Story 1	Story 2	Common structure
shows id371 id372	**declare-war id818 id819**	declare-war id818 id819
offers-exchange id371 id372 id373	sings id207 murder	decides-to-react ?x1
not-perform-service id373	**decides-to-react id142**	sets-out ?x1
negative-result id373	**sets-out id142**	wins ?x1
consumes id373 id44	**wins id142**	brings-peace ?x1
acquires id373 magical-abilities	**brings-peace id142**	arrives ?x1 ?x2
declare-war id818 id819	**arrives id142 id730**	disguised ?x1
dispatches id189 id373	**disguised id142**	unrecognised ?x1
tells id189 id373 past-misfortune	**unrecognised id142**	claims id672 won id818
decides-to-react id373	**claims id672 won id818**	sets ?x3 ?x1
sets-out id373	**sets id165 id142**	involves difficult-task ?x4
wins id373	**involves difficult-task strength**	solve ?x1 difficult-task
brings-peace id373	**solve id142 difficult-task**	before dead-line
arrives id373 id728	**before dead-line**	returns ?x1
disguised id373	**returns id142**	arrives ?x1 id730
unrecognised id373	**arrives id142 id730**	disguised ?x1
claims id672 won id818	**disguised id142**	unrecognised ?x1
sets id161 id373	**unrecognised id142**	claims id672 won id818
involves difficult-task kissing	**claims id672 won id818**	exposed id672
marked id373	**exposed id672**	not-solve id672 difficult-task
solve id373 difficult-task	**not-solve id672 difficult-task**	
before dead-line	new-physical-appearance id142	
returns id373	punished id818	
arrives id373 id730	tied-to id818 horse-tail	
disguised id373		
unrecognised id373		
claims id672 won id818		
exposed id672		
not-solve id672 difficult-task		

Table 1: Table of events in each of the stories and the shared set of events.

3 A Callibration Exercise for Story Similarity

Although there are many possible representations for stories and many different metrics have been considered for story similarity, the present effort has been focused on a particular representation format as used by an existing story generator, and a specific metric that allows automatic computation. These choices were circumstantial on ease of access and are not considered optimal, but the effort should produce valuable insights that can later be extended to other alternatives.

3.1 Story Representation in the Propper System

The Propper system [5] constitutes a computational implementation of a story generator based on Propp's description of how his morphology might be used to generate stories [13]. It produces stories as a sequence of states described in terms of predicates that hold in the state. Characters, objects or locations are represented as unique identifiers in the predicates. This representation format has been considered generic enough to allow for an initial calibration exercise, considering that other formats may easily be converted into this one.

The representation includes predicates representing narrative events and predicates describing properties of the characters that hold in particular states

of the story. These appear jointly in the stream of predicates for the story, but have been separated in the presentation of stories in this paper for clarity.

The predicates presented here result from an effort of reverse engineering of the stories that Propp describes as examples of the application of his framework to analyse existing Russian folk tales.

The first two columns of Table 1 present two examples of the stories produced by the Propper system. Predicates in this table describe actions or events in the story. Table 2 represents non-narrative facts that are true for the arguments of the actions in Table 1.

Story 1	Story 2	Common structure
hero id373	*villain id818*	*villain id818*
donor id371	*victim id819*	*victim id819*
magical-agent id372	*hero id142*	*hero ?x1*
magical-agent id44	*seeker-hero id142*	*location ?x2*
villain id818	*location id730*	*false-hero id672*
victim id819	*court id730*	*unknown ?x3*
seeker-hero id373	*groom id142*	*task-type ?x4*
dispatcher id189	*false-hero id672*	*court id730*
location id728	*location id730*	
home id728	*court id730*	
apprentice id373 artisan	*groom id142*	
false-hero id672		
location id730		
court id730		
groom id373		

Table 2: Table of characters, locations and objects in the two stories and the shared set

3.2 Human Interpretation of the Stories

In order to compare the human interpretation of the stories with an automatically extracted report, we asked a human expert to write both stories in English and compare them. It is important to mention that the expert was familiar with this type of representation based on predicates, but she had to figure out the meaning of the predicates based solely on their names.

Story 1

This story has the following main characters: a hero (373), a villain (818), and a false hero (672). In addition, a donor (371), a victim (819) and a dispatcher (189) appear as secondary characters.

The hero (373) is first offered a magical agent by a donor (371) if he performs a service. He does not perform the service but he obtains another magical agent anyway, which he consumes to acquire magical abilities.

Then, a villain (818) appears who declares war to a victim (819). The victim does not appear again.

Meanwhile, a dispatcher (189) talks about a past misfortune. The hero decides to react, sets out and wins (the war?), bringing peace with him. After that, the hero goes home, but he is disguised as the apprentice of an artisan and is not recognised. He finds a false hero (672) at home, who claims that he defeated the villain.

The hero is marked, solves a difficult task and returns to the court, this time disguised but as a groom. The false hero still claims that he defeated the villain, but he is exposed and it is known that he did not solved a difficult task.

Story 2

This story has the following main characters: a hero (142), a villain (818), and a false hero (672). In addition, a victim (819) appears only at the beginning.

The story starts with the villain (818) declaring war to the victim (819). The hero (142) decides to react, becomes a seeker hero, sets out and wins (the war?). He brings peace and arrives to the court. But he is disguised as a groom and he is not recognized.

At the court, the false hero (672) claims that he defeated the villain. Someone (165) sets the hero a difficult task that involves strength. He solves the difficult task before the deadline, and returns to the court. Again he is disguised as a groom and he is not recognized.

And again, the false hero claims that he defeated the villain. However, the false hero is exposed and does not solve a difficult task. The hero gets a new physical appearance (undisguised?), and the villain is punished being tied to a horse tail.

Next, we asked the expert to compare both stories and describe the main similarities and differences between them.

Both stories are similar in their characters and roles: a hero, a villain, and a false hero who claims to have defeated the villain.

In addition, in both stories the villain declares war to a victim, and the hero wins the war and brings peace. After that the hero returns (home or to the court) disguised (as a groom or as an apprentice), and he finds that a false hero claims to have defeated the villain. But at the end the false hero is exposed in both stories. Also, in both stories the hero makes two different journeys: one to win the war and return home/court, and one to solve a difficult task and then returning to court.

From the point of view of the differences, Story 1 involves magic. The hero tries twice to obtain a magical agent, and the second time he achieves it and gets magical abilities. However, they are not used in the story. The main difference in Story 2 is that at the end the villain is explicitly punished by being tied to a horse tail.

It is interesting to note that the first things mentioned by the expert both in the descriptions and the comparison are the characters, although in the comparison only the most important characters are mentioned, as the others are considered less important for the plot.

In addition, the descriptions are based on the most important events in the story, so not all events are considered equally important. The comparison also shows that there is a high similarity between both stories in terms of characters and some of the narrative arcs. For example, the hero returns in both stories but to different places and with different disguises. However, these differences (place and disguise) are not considered as important and the expert finds similarity in what is happening even when the stories are not exactly the same.

One of the main differences between the stories is that one of them involves magic, but it is not considered so important because magic is not used in the rest of the story. Finally, the differences in the endings are explicitly addressed in the comparison. This means that the end of the story is an important part of it.

3.3 Computing the Common Structure of Two Stories using Plan Refinements

A story in its more basic form can be represented as a sequence of actions, i.e., as a *plan*. There are different approaches to compute the similarity of two plans. In this paper we use the similarity measure based on plan refinements presented in [15] because it does not only provide a numerical similarity value but an explicit description of the common structure shared by both plans. This common structure can be seen as a directed graph in which each node represents an action and each directed edge represents an ordering constraint. Two actions are connected in the graph only if both actions appear in that order in the plans being compared.

Besides the actions and their order, this similarity measure also considers the action parameters and, if they are different in both plans, it is able to infer their common type according to a domain taxonomy. In this way, we are able to detect objects, characters and locations in different stories that have a different name but play the same role in the story.

The similarity measure computes this common structure performing successive refinements in the space of partial plans [7]. There are five different types of refinements that specialize a partial plan: to add a new action, to add a new ordering constraint between two existing actions, to specialize the type of a variable representing an action parameter according to a domain taxonomy, to unify two different variables, and to replace a variable with a domain constant.

The similarity measure works as follows. Let us suppose we want to compare two plans (or stories) p_1 and p_2. The similarity measure begins with an empty partial plan (a plan with no actions) that represents any possible plan and thus it is more general than p_1 and p_2. Then the partial plan is specialized using a refinement operator (adding new actions and ordering constraints or specializing the action's parameters) until we reach another partial plan that cannot be specialized anymore while being more general that both p_1 and p_2. This partial plan is the *most specific generalizer* of p_1 and p_2, $MSG(p_1, p_2)$, and represents the common structure shared by the two plans. The length of the refinement chain from the empty plan to the $MSG(p_1, p_2)$ is an indicator of how similar

the two plans are. In the same way, the length of the refinement chain from the $MSG(p_1, p_2)$ to each one of the two plans is an indicator of how much information is contained only in one of them but not in the other. The similarity value is computed as the ration between the amount of information shared and the total amount of information contained in the two plans.

The last columns of Tables 1 and 2 show the common structure computed by the similarity measure. In this case, the two stories are very similar and the inferred common graph of actions is so simple that, in fact, it can be represented as a sequence of actions. Constants representing characters, locations and objects common to both stories are kept in the common structure, and the other constants are replaced by variables with generalized types (variable names begin with '?').

The common structure of both stories could be summarized as follows. A villain declares war on a victim, what triggers the intervention of a hero that defeats him and brings peace back. Then the hero travels disguised and see how a false hero claims that he, and not the original hero, has defeated the villain. The hero leaves, solves a difficult task before some deadline, and comes back disguised. The false hero is exposed in court because he was not able to solve the difficult task.

4 Discussion

There are a number of issues that the similarity metric considered here does not take into account.

First, the point in the story in which a particular sequence of actions takes place may lead to different results. A marriage at the start of the story sets the scene for later actions, but at the end of the story it usually acts as a reward for the efforts of some character. This influence of context is not considered in the metric that has been described.

Second, some events are more significant than others. The presence of a murder in a given story is more significant than that of more mundane events such as setting off on a journey. This aspect might be captured by some kind of weighting of the importance of specific events. The described metric does not allow for this type of behaviour.

The judgment expressed by the human placed considerable emphasis on the relative importance of the elements that appear in the stories. Characters are mentioned first, then specific actions. In both cases, a certain degree of abstraction is applied to identify conceptual similarity even between instances that are different. This suggests that taxonomical reasoning might be a useful tool for assessing similarity and that, as expected, abstraction is fundamental in story similarity.

These two aspects suggest that automatic story comparison needs to address *lifting* between different levels of abstraction to be able to match those features that humans are able to match. It also seems that the abstract matching at different levels is a fundamental cognitive tool for comparing stories in humans.

This conclusion relates to the approach in [9] of considering similarity between stories in terms of a shared summary, but extended to summarisation with an important degree of abstraction. The work in [11], by virtue of being based on description logic ontologies, does include the possibility of taxonomical reasoning being applied in the process of measuring similarity. It is clear that this particular approach should be explored in more detail in future work.

The version of the Propper system that has been employed here provides only limited description of the characters. The descriptions considered are restricted to specification of the roles played in the narrative by particular characters, and a number of properties of particular arguments that are relevant for the correct chaining of later actions with their context of occurrence via their set of preconditions.

An important problem from the point of view of assessing the novelty of creative processes is the need to consider an existing set of artifacts as a reference. Generated artifacts are only novel if they are not similar to existing ones. However, from a computational point of view, the approach of keeping a record of all existing artifacts of a given type, and computing the similarity of any newly generated artifacts with this set is not practical [4]. Indexing solutions may be used to improve efficiency, but even so, solutions based on some level of abstraction, away from specific instances and addressing more generic characterisations of the artifacts (in this particular case, stories) would prove more practical in this context. Conformance or departure from Concepts such as conventional endings, genre conventions, or character stereotypes may play a fundamental role in assessing the novelty of stories beyond sequences of actions.

Overall, it seems that there are a number of aspects of stories that are relevant when attempting to establish similarity between two instances of story. Just how many such aspects should be included in a particular implementation as a similarity metric may depend substantially on the purpose for which it is intended. In the particular case of similarity metrics employed for case-based reasoning, the choice of which aspects of similarity to model should be guided by the particular aspects of the case that will be reused. If the cases are intended to provide story structure, the similarity should focus on story structure. If the cases are intended to inform decisions on the set of characters to employ, the similarity should focus on the set of characters. In relation to the point raised above concerning abstraction, it is important to note that focusing on particular aspects of story similarity may require specific types of abstraction to implement the described lifting operation. Where similarity metrics are used for evaluating novelty in Computational Creativity settings, their use is much broader and it becomes more difficult to focus on particular aspects. Nevetheless, as it is very important to consider issues of efficiency, abstraction as means of reducing the range of attributes that need to be compared will clearly play a fundamental role in practical implementations.

5 Conclusions and Future Work

The present work describes a process by which a computational system for computing the similarity between narrative structures is compared and calibrated against human judgment.

A number of issues considered by the human judge but not covered by the system have been discovered. These should be considered as a check list for the consideration of alternative metrics, and possibly as driving guidelines for the development of more elaborate metrics specific to the assessment of story similarity.

The work described in this paper has addressed sequential single narrative threads. More complex narratives usually involve parallel story lines which merge or split at several points in the overall narrative. Whether the current metrics are valid for comparing similarity between this kind of narratives or not is yet an open question. Additionally, the use of different structures for stories also opens a new path, namely the application of the current process to stories that, while outputting an equivalent format, are generated by other story generation systems, probably conveying different semantics in the sequence of events, and possibly richer relations between characters.

From this point of view, more recent versions of the Propper system [6] address specifically the description of characters as they occur in the story, and they should be explored in further work to extend the metric for similarity to consider differences between the characters of two stories. For that work, it may be necessary to focus on differences between characters fulfilling equivalent narrative roles in the different stories.

State is also fundamental in narrative composition and analysis. Narrative understanding of statements like "John squashed the spider" heavily depend on the relation between John and the spider (was it his mascot?). This kind of information must be taken into account in a general model of story similarity.

In all cases, further research must look into more metrics for story comparison and employ more experts to analyse how humans evaluate narratives. Following the intuition that we, as humans, perform a complex set of comparisons for evaluating similarity at different levels can lead to the discovery of plausible metrics and plausible aggregation methods into one single judgment.

References

1. Fay, M.: Story comparison via simultaneous matching and alignment. In: Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC'2012). Istambul, Turkey (2012)
2. Fisseni, B., Lowe, B.: Which dimensions of narratives are relevant for human judgments of story equivalence? In: Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC'2012). Istambul, Turkey (2012)
3. Fisseni, B., Lowe, B.: Event mappings for comparing formal frameworks of narratives. Logique et Analyse (57) (2014)

4. Gervás, P.: Dynamic inspiring sets for sustained novelty in poetry generation. In: Second International Conference on Computational Creativity. México City, México (2011)
5. Gervás, P.: Propp's morphology of the folk tale as a grammar for generation. In: Workshop on Computational Models of Narrative, a satellite workshop of CogSci 2013: The 35th meeting of the Cognitive Science Society. Universität Hamburg Hamburg, Germany (2013)
6. Gervás, P.: Computational drafting of plot structures for russian folk tales. Cognitive Computation (2015)
7. Kambhampati, S., Knoblock, C.A., Yang, Q.: Planning as refinement search: A unified framework for evaluating design tradeoffs in partial-order planning. Artificial Intelligence 76(1), 167–238 (1995)
8. Krakauer, C., Winston, P.: Story retrieval and comparison using concept patterns. In: Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC'2012). Istambul, Turkey (2012)
9. Kypridemou, E., Michael, L.: Narrative similarity as common summary: Evaluation of behavioral and computational aspects. LLC 29(4), 532–560 (2014), http://dx.doi.org/10.1093/llc/fqu046
10. Michael, L.: Similarity of narratives. In: Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC'2012). Istambul, Turkey (2012)
11. Peinado, F., Francisco, V., Hervás, R., Gervás, P.: Assessing the novelty of computer-generated narratives using empirical metrics. MINDS AND MACHINES 20(4), 588 (2010)
12. Pérez y Pérez, R., Ortiz, O., Luna, W.A., Negrete, S., Pealoza, E., Castellanos, V., vila, R.: A system for evaluating novelty in computer generated narratives. In: Proceedings of the Second International Conference on Computational Creativity. pp. 63–68. Mxico City, Mxico (2011)
13. Propp, V.: Morphology of the Folk Tale. Akademija, Leningrad (1928)
14. Ritchie, G.: Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds & Machines 17, 67–99 (2007)
15. Sánchez-Ruiz, A.A., Ontañón, S.: Least common subsumer trees for plan retrieval. In: Case-Based Reasoning Research and Development - 22nd International Conference, ICCBR 2014, Cork, Ireland, September 29, 2014 - October 1, 2014. Proceedings. pp. 405–419 (2014)