Extending and Evaluating a Platform for Story Understanding

David K. Elson and **Kathleen R. McKeown**
Columbia University Department of Computer Science
New York City

Abstract

We summarize recent developments in our platform for symbolically representing and reasoning over human narratives. The expressive range of the system is bolstered by the infusion of a large library of knowledge frames, including verbs, adjectives, nouns and adverbs, from external linguistic resources. Extensions to the model itself include alternate timelines (imagined states for goals, plans, beliefs and other modalities), hypotheticals, modifiers and connections between instantiated frames such as causality. We describe a corpus collection experiment that evaluates the usability of the graphical encoding interface, and measure the inter-annotator agreement yielded by our novel representation and tool.

Introduction

In our paper at the previous Symposium on Intelligent Narrative Technologies (Elson and McKeown 2007), we presented SCHEHERAZADE [1] as a platform for narrative intelligence that formally represents stories. Our goal is to offer to the community a versatile tool for authoring, encoding and representing stories symbolically; its formal model enables feature extraction and machine learning for story understanding. This can apply to tasks such as story classification on the thematic dimension (beyond the lexical and syntactic features of the text), genre analysis and authoring support (providing feedback about a new story based on insights from empirically derived models of story structure).

One key feature of this approach is the separation of *narrative semantics* – temporal order, argument control for actions, and other aspects common to all stories – from domain knowledge (the actions and objects possible in the story-world) and the content of an individual story. Though this approach makes deep inference about the story-world difficult (the system cannot, for example, compute the actions necessary to achieve a goal state), it allows us to model a wider range of narratives than other approaches such as plan-based representations. Its expressive range is as large as the set of domain knowledge modeled for the story-world.

Our current project is to build a novel type of corpus: a collection of *story graphs* that combine linguistic and semantic features. Specifically, the graphs consist of nodes and arcs representing core narrative elements, such as a story *state* and the actions which occur during that state. The model represents actions as *predicates*, which include a verb *frame* and the arguments filling the frame's thematic roles. For example, the predicate `depart(king)` invokes the *depart* frame with *king* as the agent. Objects such as *king* are themselves frames to be instantiated; one can express that a generic king or some particular king is departing.

We have written a graphical interface that facilitates the building of story graphs through animated timelines, predicate construction forms and a natural-language generation component that expresses predicates back into prose. In other words, the user does not see frames and predicates themselves, but only clauses and sentences such as *The king departed.* This method hides the semantic underpinnings of the model to allow non-technical users to construct elaborate story graphs based on existing texts or their own creativity.

Previously, our results indicated that the model, along with its interface and set of domain knowledge, were too limited for us collect a corpus of detailed story graphs encoding the fables attributed to Aesop. In the following sections, we describe significant extensions to the model and the infusion of over 100,000 knowledge frames to enhance the expressiveness of SCHEHERAZADE. We also describe an encoding experiment designed to evaluate the platform's usability and formality.

World knowledge infusion

One clear bottleneck for the expressive range of a representation and encoding tool is the authoring of action and object types to be made available for use in a story. Previously, SCHEHERAZADE included a small set of hand-crafted conditions (which are statives, e.g., *a character desires an action*), actions (which are state-changing, such as *a prop falls*) and object classes. To make the system viable for a range of domains and narrative scenarios, we turned to external linguistic resources to supply a wealth of missing frames.

For objects, WordNet (Fellbaum 1998) was the natural choice. This well-established lexicon features thousands of words organized into *synsets* with the same meaning. One synset, for example, includes the nouns *meadow* and *hayfield*. Synsets are organized into hypernym trees, with each synset related to more and less specific synsets. As our

[1] `http://www.cs.columbia.edu/~delson/`

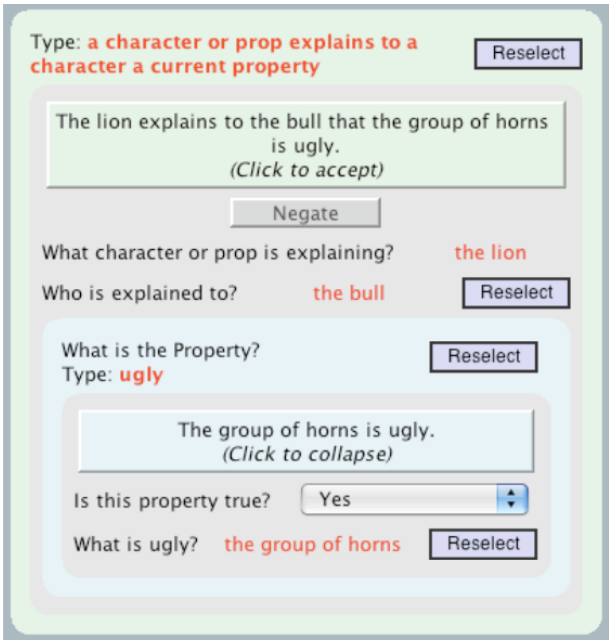

Figure 1: The section of the semantic encoding interface allowing annotators to create predicates from a library of world-knowledge frames. Here, the annotator is modeling a sentence from Aesop's fable *The Wily Lion* by nesting two predicates together, one indicating that the other is dialogue.

model also supports hierarchical relationships, we needed only to decide which subtrees of the root noun synset (*entity*) were to be imported for each SCHEHERAZADE object type. For example, to populate our list of available character types, we adapted the *organism* subtree, allowing users to write about thousands of animal species or roles such as *traveler*. Overall, we imported 11 subtrees for prop types (covering about 47,000 nouns), 3 subtrees for character types (30,000 nouns) and 10 subtrees for location types (15,000 nouns). WordNet's adjectives, meanwhile, serve as conditions that describe people or things: *The king was mighty*.

While WordNet provides a hypernym tree for verbs as well, there is limited information about each verb's semantic arguments. We turned to VerbNet (Kipper et al. 2006), the largest online verb lexicon currently available for English, for these details. Each verb is annotated with thematic roles (arguments) and their selectional restrictions, as well as syntactic frames for the various sentence constructions in which the verb and its roles might appear (which we adapt to serve as plans for the system's generation component).

Users of the tool can therefore construct elaborate predicates without having to model the *types* of actions or objects available in the story-world. For example, if a user searches for "guide" in the graphical interface, she sees a list of available frames from which to choose: *A character guides a character*, *A character guides a prop*, and so on. Upon selecting the desired frame, the interface prompts her to supply each argument required by the verb frame by asking specific questions based on the selectional restrictions. For instance, it will handle an *Agent* thematic role by asking the user which character is doing the action from among a list of available characters. Some *Theme* roles become calls for nested action or condition predicates, such as in dialogue verbs such as *explain* (see Figure 1).

All told, these resources have supplied more than 92,000 object frames, 27,000 action frames and 14,000 condition frames, greatly enhancing the expressive power of the platform. We also turned to an additional lexicon (namely, COMLEX (Macleod, Grishman, and Meyers 1994)) to supply plural forms and irregular conjugations for the generation component.

Model extensions

Among the additions to the underlying model we have implemented are modifiers, connectives between predicates, alternate modalities and the ability to negate actions (*he didn't speak*). In this section, we describe the major changes.

In the previous version of SCHEHERAZADE, conditions were used to modify nouns, but there was no way to modify condition and action predicates themselves. We now model modifiers as a class of predicates, represented as nodes in the conceptual graph attached with arcs to actions, conditions or other modifiers. There are two classes of modifiers: adverbials (e.g., *slowly*) and connectives, which link multiple actions in the story together with relationships such as causality. In other words, users can now express that one predicate takes place because of, or despite, another predicate elsewhere in the story. In the internal representation, this is modeled with an arc between the two nodes in the story graph, in the manner of QUEST (Graesser, Lang, and Roberts 1991). The grammar of our generation module inserts adverbs and subordinate clauses appropriately, such as *The man limped because he was ill.* For an action that spans multiple states, users can assign a modifier to the entire action or limit the scope to its cessation: *The man gradually stopped limping.* We adapted thousands of adverbs from WordNet to serve as modifiers.

Furthermore, we have recently introduced support for *alternate timelines* in the encoding interface. An alternate timeline is a separate scope for one or more predicates, which are then attached as a set to the main timeline under some temporal relationship and modality. This capability is useful for goals, plans, beliefs, hopes and fears, where actions are imagined and do not necessarily occur. We also use them to model repeated actions (e.g., *he cooked eggs and ate them daily*), actions in the past and future (*he had eaten the cookies*), and possible states (*it was possible that Donald would eat the cookies, then get sick*). There are special condition frames such as *A character desires a timeline* that a user can select to attach an alternate timeline.

Every action and condition can be set to be a conditional rather than actual occurrence within the scope of its timeline. Other actions and conditions in the same timeline are then considered to be contingent on the fulfillment of the conditional predicates. The generation component chooses the correct tense and aspect based on conditionality and the semantics of what happens when: *Donald hoped that – if it were to begin raining – he would have finished playing.*

Collection and evaluation

Our purpose for SCHEHERAZADE is to build a corpus of semantically encoded stories from which we can perform machine learning over the thematic dimension of narrative. One test of the suitability of the tool is to measure inter-annotator agreement – the degree to which different annotators agree on the appropriate semantic model for a clause or sentence. As the model becomes more expressive, users may find different representations for a clause that all convey the idea of the original text; this divergence can affect machine understanding. In this section, we discuss a collection experiment that measures similarities between story graphs by different annotators based on the same fables.

Collection

We recruited 9 annotators from both our department and other undergraduate schools. After approximately 20 minutes of training, we assigned each annotator one or two stories to encode, with the goal of bringing the "reconstructed story" (a reading of the entire graph by the generation component) as close as possible to the original text. For those annotators who completed two encodings, we varied the ordinal position of each fable.

We collected 17 encodings, grouped into 4 parallel encodings for each of 4 separate fables (with one fable getting 5 encodings). On average, excluding predicates in alternate timelines such as goals, each encoding consisted of 16.7 predicates including 11.5 actions and 5.2 conditions. The annotators wrote an average of 1 predicate for each 8 words in the source text, which varied from 80 words (11.5 average predicates) to 175 words (20.6 average predicates). We found that while the annotators who encoded two stories tended to take slightly less time on the second story, there was no correlation between the length of an encoding and its ordinality (first encodings and second encodings, on average, exactly matched the mean length of all encodings for their respective fables). This suggests that while annotators became more familiar with the encoding tool over time, all of their encodings were equally thorough. See Figure 2 for an example encoding, as expressed from the story graph by the generation component.

We asked the annotators to complete a survey after finishing each encoding. The survey asked annotators to rate the tool's usability and to list aspects of the story that were not encodable or were encoded oddly. On a Likert scale from 1 (most difficult) to 5 (easiest), the average rating was 2.95 among first encodings and 3.2 among second encodings; though the span is not statistically significant, their comments indicate that the task of semantic encoding is challenging at first but becomes easier with practice. The most frequently cited deficiencies in the model were abstract nouns such as *idea* and *kindness*, though we believe better learning results will occur when users unpack such abstractions into concrete actions and conditions. We have since implemented other interface and model improvements requested by the annotators, and we continue to improve the system iteratively based on ongoing collection experiments.

Fable	1	2	3	4
Parallel encodings	4	5	4	4
κ maximum	.44	.45	.25	.34
κ average	.33	.27	.19	.26
κ st. dev.	.09	.08	.05	.06

Table 1: Pairwise agreement scores for encodings of *The Fox and The Grapes*, *The Mouse and The Lion*, *The Serpent and The Eagle* and *The Wily Lion*, respectively.

Corpus Analysis

Our basic measure for comparing two encodings is the similarity s between some two predicates a and b. Similar to prior work (Budanitsky and Hirst 2001), we find the most specific common ancestor for two predicate types in the hypernym tree (that is, the lowest common ancestor), and assign a score relative to the path length to that ancestor. We also consider the overlap between the arguments (i.e., the attributes, as previously suggested in (Tversky 1977)), by recursively scoring each argument predicate. Both measures contribute evenly to a final score between 0 and 1, by the formula

$$s(a,b) = \frac{\frac{1}{1+h(a,b)} + \frac{\sum_{i=1}^{r(a,b)} s(p(a,b,i))}{r(a,b)}}{2} \quad (1)$$

where

- $h(a,b)$ is the average path length from the two predicate types to their common ancestor in the hypernym tree, or ∞ if their only common ancestor is the root type
- $r(a,b)$ is the size of the union of the thematic roles covered among the arguments to both predicates, and
- $p(a,b,i)$ retrieves the nested predicates which serve as arguments in a and b for some thematic role i.

For scoring the similarity between two entire encodings, we find the sum of the pairwise similarity scores among all corresponding pairs of predicates (those modeling the same concept), then normalize the sum for the sizes of the encodings. We adapted the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), devised for aligning DNA sequences, to automatically determine the most likely mapping between predicates in one encoding and their counterpart predicates in the other encoding.

Results

In general, we would like the similarities between parallel encodings of the same story to be high. We use Cohen's kappa (Cohen 1960) to measure the agreement between annotators with respect to a baseline of chance agreement. We determine chance agreement by generating two random narratives of similar length and calculating their pairwise alignment score. The kappa scores for our four encoded fables are given in Table 1.

These agreement levels indicate a fair number of analogous predicates, but significant differences as well. There are several factors introducing variability in encodings of the same story:

A Lion watched a fat Bull feeding in a meadow, and his mouth watered when he thought of the royal feast he would make, but he did not dare to attack him, for he was afraid of his sharp horns.
Hunger, however, presently compelled him to do something: and as the use of force did not promise success, he determined to resort to artifice.
Going up to the Bull in friendly fashion, he said to him, "I cannot help saying how much I admire your magnificent figure. What a fine head! What powerful shoulders and thighs! But, my dear friend, what in the world makes you wear those ugly horns? You must find them as awkward as they are unsightly. Believe me, you would do much better without them."
The Bull was foolish enough to be persuaded by this flattery to have his horns cut off; and, having now lost his only means of defense, fell an easy prey to the Lion.

There once was a fat and horned bull and a hungry lion. The lion was watching the bull. The bull was eating from a field.
The lion wanted to eat the bull because the bull was fat and didn't attack him because the bull was horned.
The lion began to be cunning because he was hungry.
The lion approached the bull and began to be friendly.
The lion explained to the bull that the bull was extremely attractive and explained to him that a group of horns was ugly.
The bull began to be foolish, believed the lion, removed the group of horns and began to be hornless.
The lion ate the bull because the bull was hornless.

Figure 2: *The Wily Lion* (top) and a subject's encoding.

1. The high expressivity of the model allows roughly the same idea to be expressed correctly in multiple ways. For example, one annotator indicated "The lion is asleep" as a condition, while another chose "The lion was sleeping" as a progressive action. The two frames share no ancestor.
2. Some paraphrases are less clear-cut and may, in fact, represent different subjective interpretations of the story. For example, one annotator summarized a description of dialogue in *The Wily Lion* with "flatters" where another modeled predicates for the specific compliments. Did the second annotator disagree that the dialogue was flattery, or simply choose to create more detailed predicates?
3. Some differences come from sampling error, i.e., noise introduced by the tool itself during the collection process. Some annotators may have felt more comfortable with semantic modeling than others, and were thus more able to explore certain corners of the tool such as conditionality.

We are currently working to determine the balance between these three sources of variation. We will work to minimize the first and the third sources with methods to normalize minor paraphrases and improve the user interface based on annotator feedback. The second source of variation, though, is a valuable insight to be preserved. We intend a story graph to encode the annotator's inference of the semantic underpinnings of the story based on the text. Similarly, prior work has examined the differences between subjective interpretations of a story to study the mind's representations (McKoon and Ratcliff 1992); even a single individual will tell different versions of a story over time (Passonneau, Goodkind, and Levy 2007). In this manner, the platform can be used to measure not only the thematic differences between stories, but differences in story understanding between individuals and over time.

Conclusion

With a boost from external sources of world knowledge, significant extensions to the model and a more flexible interface, our system has matured to the point where collecting a corpus of narrative encodings is viable, even from users who are not versed in linguistics or computer science. While more work remains for supporting a wider range of assertions and normalizing paraphrases, SCHEHERAZADE offers a robust platform for creating a novel type of linguistic-semantic corpus.

References

Budanitsky, A., and Hirst, G. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NACCL 2001 Workshop: on WordNet and other lexical resources*, 29–34.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Elson, D. K., and McKeown, K. R. 2007. A platform for symbolically encoding human narratives. In *Proceedings of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies*.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Graesser, A.; Lang, K.; and Roberts, R. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General* 120:254–277.

Kipper, K.; Korhonen, A.; Ryant, N.; and Palmer, M. 2006. Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*.

Macleod, C.; Grishman, R.; and Meyers, A. 1994. Creating a common syntactic dictionary of english. In *Proceedings of SNLR: International Workshop on Sharable Natural Language Resources,*.

McKoon, G., and Ratcliff, R. 1992. Inference during reading. *Psychological Review* 99(3):440–466.

Needleman, S. B., and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3):443–453.

Passonneau, R.; Goodkind, A.; and Levy, E. 2007. Annotation of children's oral narrations: Modeling emergent narrative skills for computational applications. In *Proceedings of the 20th Annual Meeting of the Florida Artificial Intelligence Research Society (FLAIRS-20)*.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84(4):327–352.