Corpus Annotation in Service of Intelligent Narrative Technologies

Mark Alan Finlayson
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Room 32-258
Cambridge, MA 02138 USA
markaf@mit.edu

Abstract

Annotated corpora have stimulated great advances in the language sciences. The time is ripe to bring that same stimulation, and consequent benefits, to computational approaches to narrative. I describe an effort to construct a corpus of semantically annotated stories. I outline the structure of the corpus, a structure which colloquially can be described as a "handful of handfuls." One handful of the corpus has already been constructed, viz., 18k words of Russian folktales. There are two handfuls under construction: legal cases focused on the area of probable cause, and stories from Islamist Extremist Jihadists. Four more handfuls are being planned: folktales from Chinese, English, and a West Asian culture, and stories of international conventional and cyber conflicts. There are numerous additional handfuls under discussion. The main focus of the corpus so far has been on textual materials that are annotated for their *surface semantics* using conventional annotation tools and techniques; nonetheless, there are numerous novel dimensions along which the corpus might grow and become useful for different communities. In particular I propose for discussion the outlines of a few novel sources, annotation schemes, and collection methodologies that could potentially make the corpus of great use to the interactive narrative or narrative generation communities.

Annotated corpora have stimulated great advances in the language sciences. When easily available and widely used, they provide numerous advantages, including reducing research costs and consolidating effort by providing pre-curated and annotated data on which everyone can build, forming training data for machine learning, providing a gold standard against which a automatic methods can be evaluated, and allowing comparison between studies done with different methods and at different institutions. To illustrate with just a few examples (of many), the Penn Treebank streamlined and focused work in developing automatic syntactic parsers (Marcus, Marcinkiewicz, and Santorini 1993). When the Treebank was released, statistical syntactic parsing technology was in its infancy; now parsers are robust, have wide-coverage, and serve as the foundation for a variety of other natural language processing tasks. PropBank, another example, did for the semantic role labeling what the Treebank did for syntactic parsing (Palmer, Kingsbury, and Gildea 2005). The ability of annotated corpora to stimulate advances has even gained its own meeting series: SemEval (Kilgarriff and Rosenzweig 2000) focuses on tasks built around annotated corpora, and has led to significant advances in areas such as named entity recognition, word sense disambiguation, co-reference clustering, and event identification.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Computational approaches to narrative lack a resource analogous to the Penn Treebank that would provide the same benefits. I propose that the time is ripe for construction of such a corpus. For lack of a better name, I will refer to it as the *StoryBank*. In the following section I describe the StoryBank's proposed gross structure, a structure which crystalized out of discussions that were held at the AAAI 2010 Fall Symposium on Computational Models of Narrative (Finlayson et al. 2010; Finlayson 2011) held in Arlington, Virginia. The structure is extensible and designed to allow for exploration and identification of useful approaches. I also describe a set of representations, and associated tool, that will form the core set of annotations for the corpus. I also describe the portions of the StoryBank that have recently been completed, the portions that are underway, and the portions that have been mapped out but not yet begun.

In the second section of the paper, I discuss the potential utility of the StoryBank for domains of special interest to the Intelligent Narrative Technologies (INT) community: interactive narrative and narrative generation. While it is still speculative that the StoryBank would be of use in those domains, I propose a number of novel types of data that might be incorporated, collection methodologies that might be explored, and annotation schemes that could be useful. It is my hope to entice interest in the StoryBank within the INT community, and engage their participation in the StoryBank's design, construction, and use.

The StoryBank Corpus

There are three major aspects of the corpus on which there has been some agreement. First, the gross structure of the StoryBank. Second, the identity of a fraction of its contents: numerous other proposals are on the table and ripe for discussion. Third, the annotation schemes and tools that form the common core for all the parts of the corpus.

Gross StoryBank Structure

A great deal of time at the AAAI 2010 Fall Symposium on Computational Models of Narrative (hereafter: CMN-2) was devoted to discussing the potential of building a resource like the StoryBank. The meeting was attended by over 40 researchers from over 10 countries in the Americas, Europe, and Asia; they brought expertise from many areas, including commonsense reasoning, formal logic, natural language processing, language generation, representational formalisms, analogical reasoning, legal reasoning, argumentation theory, geospatial narratives, interactive narrative technologies, cognitive science, cognitive narratology, linguistics, discourse analysis, cognitive psychology, anthropology, sociology, and philosophy. The main worry of such a diverse group was that the corpus would be of use only to a small subset of the community if it focused exclusively on a particular type of narrative, topic of interest, or annotation scheme.

Thus, perhaps the most valuable idea[1] to come of the discussion was that of structuring the corpus as a "handful of handfuls": instead of collecting a monolithic corpus with a vast number of words representing a single narrative type and topic, with a single annotation scheme, the corpus could be structured around small batches, or 'handfuls,' that each clearly serve some research area. Each handful might potentially contain annotations specific to its purpose (although with significant overlap between them). A handful of these handfuls could be collected and integrated, and interaction within the community (and observation of the resultant work) could determine what was useful and stimulating and should be expanded and replicated in other handfuls, where new additions were needed and in what new directions, and what was not working and perhaps should not be pursued. This would lead to a diverse corpus that covered a large range of types of narrative, while still maintaining flexibility without wasting too much money on dead ends. In the end the loose target agreed upon for the first stage of the StoryBank was 10-15 handfuls of 25-50k annotated words each across 15-50 stories, each handful also containing additional unannotated text, in the same domain and type, on the order of five to ten times the amount annotated.

StoryBank Contents

Of the StoryBank handfuls considered so far, one has been completed, two are in progress, and four are in the planning stages.

Currently in hand are 18k words of Russian folktales, comprising 15 stories ranging in length from 646 to 1,934 words each, drawn from English translations of Alexandr Afanas'ev's collection of Russian folktales (Guterman and Afanas'ev 1945, for example). This collection was the pilot handful on which the development of the core annotation scheme and annotation tool were tested and validated (see next section).

Two handfuls are in progress. The first is a set of U.S. District and Supreme Court cases in the domain of Probable Cause (U.S. Fourth Amendment constitutional law). We have collected and curated a set of 143 cases which have been edited to retain the "story" of the case while removing extraneous text; that is, what is retained is the statement of facts of who, what, where, and when, and the final judgement. In their edited state, these cases comprise 125k words, and range from 179 to 2307 words each. We are in the process of selecting a subset of these cases, on the order of 25k-50k words, for annotation in the StoryBank core annotation scheme (described in the next section).

The second handful underway is 40k words of stories in the domain of Islamist Extremist Jihadism, being collected in collaboration with Steven Corman and Jeffry Halverson at Arizona State University (Halverson, Goodall, and Corman 2011). These include Al Qaeda promotional stories (approximately 10k words), stories from Hadith (commentary on scripture) that are preferentially used by Jihadists to justify their philosophy (about 15k words), and battle reports and autobiographical stories drawn from Jihadist websites and other sources (approximately 15k words)[2]. These are also being prepared for annotation in the StoryBank core annotation scheme.

In the planning stage are four more handfuls. The first three are collections of folktales, akin to the Russian handful, but focused on different cultures: Chinese, English, and a West Asian culture to be determined, likely Iraqi, Iranian, or Afghan. In addition to these folktale handfuls, we are planning a handful focused around stories of international conflict, to include descriptions of both conventional (Ciment 1999) and cyber wars. Ideas for numerous other handfuls have been discussed, especially at the first workshop on Computational Models of Narrative (Finlayson, Richards, and Winston 2010). These include, for example, stories capturing legal argumentation (Bex et al. 2010); stories focused on commonsense (Mueller 2007); or visual stories such as comics (Cohn 2010).

Core StoryBank Annotation Scheme

Computational linguistics has provided us with a large suite of representations that are sophisticated enough, and vetted enough, to reliably capture a large fraction of the semantics that any typical reader will glean from reasonably complex narratives. We have fixed a set of 15 core annotations that will serve as a common interlingua between all the different handfuls of the StoryBank, the conjunction of which gives fairly reasonable cover of the basic meaning of a narrative, which I call the *surface semantics* of a text. These representations include syntax (tokens, multi-word expressions, sentences, parts of speech, lemmas, CFG parses), discourse (referring expression, coreference groups), and semantics (wordnet senses, referent properties, referent relationships, semantic roles, as well as events, temporal expressions, and temporal order).

These annotations will come with an annotation tool, called the Story Workbench (Finlayson 2008) that can be used to view and modify the annotations on the texts, and will include a Java API for programmatically interacting

[1] The credit for this must go to Livia Polanyi.

[2] The jihadi internet stories are subject to distribution restrictions, so it is as yet unclear if they will be included in a public release.

with the data. The Story Workbench is a fully functional tool, having been used so far by more than 12 different annotators to annotate over 100k words of text in over 17 different representations. It is designed to be extensible, and so additional annotation schemes may be added in the future as the corpus evolves.

Augmenting the StoryBank for INT

The primary goal of the StoryBank is to be as useful as possible to as many computational narrative researchers as possible. In this section I consider what might be done to make the StoryBank more useful those working in interactive narrative and narrative generation: two topics of great interest to many INT participants. Fair warning: this section is speculative and full of untested ideas. Some are, no doubt, impractical; some perhaps not useful. But my aim is to stimulate discussion and thought. I will draw on the productive discussion at CMN-2 for examples and ideas.

Sources

As noted, the StoryBank first stage target is 10-15 handfuls of stories, each focused on different topics, types, and annotations. Above I outlined seven handfuls; therefore there is plenty of room to add more. Here's one. At CMN-2 one of the invited speakers was a professional storyteller, Loren Niemi. He gave a storytelling performance during which he told multiple versions of the 'same' narrative, namely, *Little Red Riding Hood*. He began with simple early versions of the story; then he re-told the Brothers Grimm version; he finished with a modernized version that might be described 'narratologically sophisticated.' The different versions had emphases ranging from warnings to children, to entertainment, to reflections on human sexuality. Multiple versions of the same story: this is a topic near the heart of narrative generation. I propose that one handful of the corpus focus on a small number of seed stories, each with multiple retellings. It is easy to find, for folktales, multiple versions of the same story: pick, for instance, any tale re-told by the Brothers Grimm. If these are not suitable for some reason, it would be not much harder to recruit professional storytellers to re-tell the seed stories: indeed, this might be preferred, for it would allow data to be collected on the narrator and the audience (see next two sections).

The possible studies that could be carried out on a such a collection are numerous. Determining and annotating the dimensions along which the related stories vary would be a start; building on that, these dimensions could be used to design, guide, or train narrative generation systems. Perhaps the handful could be used to devise what would become a standard metric for narrative generation systems: a system to be tested would be asked to re-tell each seed story, perhaps multiple times, along the dimensions previously mentioned, followed by a score that compares the two sets of narratives (generated vs. corpus). One could even imagine, for a large enough corpus (probably no longer just a handful), using the stories to train statistical narrative generation systems, or populate the case-base of CBR-oriented systems.

An analogous collection of interactive narratives might be more difficult, but still potentially as useful. Perhaps we could collect transcripts of natural interactive narratives, that is, those both controlled and played by people, e.g., paper-and-pencil role playing games. Recording digital role-playing games has also shown to be feasible, e.g., instrumenting the World of Warcraft or the Second Life virtual environment.

Annotation Schemes

What annotations might serve interactive narrative or narrative generation? Here are four that spring to mind.

Plans and Goals Plans and goals are relevant not only to INT topics, but also to many other areas of narrative research. There is certainly quite a bit of work on plan and goal representations from other areas of Artificial Intelligence. Perhaps a new representation, based on previous formalisms for plans and goals, could be developed for annotating the plans and goals of characters in a narrative. For handfuls that are recordings of storytelling performances, it would be ideal if such a representation could also be used to mark the plans and goals of the *narrator* viz-a-viz the *audience*.

Emotion There has been quite a bit of intriguing work on constructing classification schemes for emotion (Ortony, Clore, and Collins 1988, for example). There has not been much, however, on actually annotating emotions in text, and especially not in stories[3]. If we had the emotions of the characters, this is information that would be of great use to producing interactive versions of narratives in the story, or producing new re-tellings.

Of course, if we can annotate the emotions of the characters, we should consider annotating the emotions of the audience and narrator. This could range from a one-word summary of the emotional impact of the performance as a whole, to a more detailed markup of individual scenes, plot points, paragraphs, sentences, phrases, or even words (more on this in the next section). If we had 50k words of stories annotated with the emotions of everyone involved, imagine what interactive narrative systems could do with that!

Interpretive Meaning This leads me up a level. Understanding what a story means to an audience, above the surface semantics, is highly relevant to interactive and generative narrative. During the CMN-2 presentations, Charlotte Linde (Linde 2010) noted that even the simplest narratives about the simplest events can hold different meanings for different people, even different people in the same culture. Is there a way to capture these multiple meanings of a story? Like the range of possible emotion representations, it might be simple, such as a single sentence that can be attached to a story that summarizes an audience member's impression, or it might be detailed, picking out individual salient words, phrases, events, or properties. Such a representation, if we had one, would be immediately applicable to other domains of computational narrative research: my first thought is of cultural tales that are moral in nature – the moral of a story could be annotated in the same way.

[3] See an interesting early effort in (Francisco et al. 2010).

Data Collection Methods

Finally, the above ideas lead me to think about to how we might collect such useful data. One important idea above is capturing the intentions, interpretations, emotions, reactions, or responses of the narrator and the audience in a live storytelling performance. It would certainly not be hard to recruit a professional story teller and a willing audience; but how do you capture their state? Perhaps the conceptually easiest is to model the collection after normal sorts of deliberative annotation collection schemes: after the performance everyone sits down in front of computers and marks up a transcript.

Another idea is to instrument the participants during the narration itself. Can we capture the audience's facial expressions? Can we monitor their vitals or skin conductance to track their emotional responses? Many Communications or Marketing departments have "Audience Response Labs" where all these data and more can be collected. Similarly, "Audience Response Systems" (ARS, a.k.a. 'clickers,' 'zappers,' or 'dial testing') are becoming popular in research and education (McCarter and Caza 2009) - it would seem straightforward to adapt them for storytelling performances.

Contributions

I outlined a vision for an annotated narrative corpus, the StoryBank, structured as a 'handful of handfuls,' each handful serving a different area of interest within the computational narrative community. The handfuls will be diverse, but unified by a set of core *surface semantics* representations. I described the current status of the StoryBank: one handful is complete, two are in progress, and four are being planned. I also laid the foundation for an important discussion on how to make the StoryBank useful for research programs related to Intelligent Narrative Technologies. I aired a number of ideas regarding sources (story variants? recordings of storytelling performances?), representations (plans and goals? emotions? interpretive meaning?), and data collection methods (post-hoc markup of transcripts? video or vital sign instrumentation? ARS?). These ideas merely scratch the surface of what is possible and what would be valuable, but they lay the foundation for what I hope will be a valuable discussion at INT-4 that will shape the direction of research for years to come.

Acknowledgments

I am grateful to the participants of the AAAI 2010 Fall Symposium on Computational Models of Narrative for their excellent ideas and suggestions. That symposium was supported in part by the Defense Advanced Research Projects Agency under contract FA8750-10-1-0076. This work was funded by the Office of Naval Research under award number N00014-09-1-0597. Any opinions, findings, and conclusions or recommendations expressed here are those of the author and do not necessarily reflect the views of the Office of Naval Research.

References

Bex, F.; van Koppen, P.; Prakken, H.; and Verheij, B. 2010. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law* 18(2):123–152.

Ciment, J. 1999. *Encyclopedia of Conflicts Since World War II*. Armonk, NY: M.E. Sharpe.

Cohn, N. 2010. Japanese visual language: The structure of manga. In Johnson-Woods, T., ed., *Manga: An Anthology of Global and Cultural Perspectives*. New York: Continuum Books.

Finlayson, M. A.; Gervas, P.; Mueller, E.; Narayanan, S.; and Winston, P. H. 2010. Proceedings of the AAAI fall symposium on computational models of narrative. In *Technical Report FS-10-04*. Menlo Park, CA: AAAI Press.

Finlayson, M. A.; Richards, W.; and Winston, P. H. 2010. Computational models of narrative: Review of a workshop. *AI Magazine* 31(2):97–100.

Finlayson, M. A. 2008. Collecting semantics in the wild: The story workbench. In *Naturally Inspired Artificial Intelligence, Technical Report FS-08-06, Papers from the AAAI Fall Symposium*, 46–53. Menlo Park, CA: AAAI Press.

Finlayson, M. A. 2011. Reports of the AAAI 2010 fall symposia: Computational models of narrative. *AI Magazine* 32(1):96–97.

Francisco, V.; Hervas, R.; Peinado, F.; and Gervas, P. 2010. Emotales: creating a corpus of folk tales with emotional annotations. In *Proceedings of the 7th LREC*, 45–85.

Guterman, N., and Afanas'ev, A. 1945. *Russian Fairy Tales*. New York: Pantheon Books.

Halverson, J. R.; Goodall, H. L.; and Corman, S. R. 2011. *Master Narratives of Islamist Extremism*. New York: Palgrave Macmillan.

Kilgarriff, A., and Rosenzweig, J. 2000. Framework and results for english senseval. *Computers and the Humanities* 34(1):15–48.

Linde, C. 2010. Social issues in the understanding of narrative. In *Computational Models of Narrative: Papers from the AAAI Fall Symposium (AAAI Technical Report FS-10-04)*, 39–40. AAAI Press, Menlo Park, CA.

Marcus, M. P.; Marcinkiewicz, M. A.; and Santorini, B. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* 19(2):313–330.

McCarter, M. W., and Caza, A. 2009. Audience response systems as a data collection method in organizational research. *Journal of Management and Organization* 15(1):122–132.

Mueller, E. 2007. Modeling space and time in narratives about restaurants. *Literary and Linguistic Computing* 22(1):67–84.

Ortony, A.; Clore, G.; and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Palmer, M.; Kingsbury, P.; and Gildea, D. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–105.