Intelligent Narrative Technologies and Social Believability in Games:
Papers from the AIIDE 2015 Joint Workshop

Toward an Automated Measure of Narrative Complexity

Sarah Harmon, Arnav Jhala
University of California, Santa Cruz
1156 High Street
Santa Cruz, California 95064

Abstract

For young children, adults learning English, or individuals with language disorders, complex narratives are difficult to create and understand. While narratives can easily be assessed in terms of their lexical and syntactic difficulty, automatically measuring the level of narrative complexity is a challenging problem. We present and evaluate a preliminary system for assessing narrative complexity, which should help identify suitable texts for readers and assist in narrative skill evaluation.

Introduction

Academic, social, and cultural development are strongly tied to progress in English reading and writing skills (Mehta et al. 2005; Nation et al. 2004). However, many individuals face obstacles that hinder their development in this area. For instance, children with language disorders often struggle to understand and create narratives at the same level as their typical-development peers (Chapman et al. 1997; Fey et al. 2004; Gillam and Johnston 1992; Greenhalgh and Strong 2001; Newman and McGregor 2006; Scott and Windsor 2000).

The task of teaching bilingual students is also a steadily increasing challenge in the United states. Between 1990 and 2011 alone, the number of Limited English Proficient (LEP) individuals in the US rose by 81% (Pandya, McHugh, and Batalova 2011). To meet the demands of bilingual students, schools must seek out reading materials of varying narrative complexity at low reading levels. The process of searching for these materials is difficult, time-consuming, and often specific to the LEP individual. Teachers are often forced to rewrite texts manually to help their students learn, and those who continue to struggle after growing up often have significantly lower health literacy and poorer health outcomes (Berkman et al. 2011; Birru et al. 2004; Koo, Krass, and Aslani 2006). An automated system to assess narrative complexity, supports the automation of this process, and can help schools overcome this challenge.

Related Work

An index of narrative complexity (INC) has been proposed to specifically address the problem of finding an appropriate measure of narrative development for children and individuals with language disorders. The INC scoring system examines thirteen narrative elements: character, setting, initiating event, internal response, plan, action/attempt, complication, consequence, formulaic markers, temporal markers, causal adverbial clauses, knowledge of dialogue, and narrator evaluations (Petersen, Gillam, and Gillam 2008). Each category is associated with a weighted score which reflects its importance in terms of narrative production skills. Evaluation of the INC scoring system suggested that its assessments are consistent and strongly correlate with the Test of Narrative Language (Gillam and Pearson 2004).

Prior work in natural language processing has involved identification of several of these narrative elements. As an example, efforts have been made in text summarization to extract the most relevant features, including character mentions, lexical aspect, and setting via parsing and named entity recognition (Kazantseva and Szpakowicz 2010). Further, some existing systems have attempted to extract these features for use in narrative understanding. For instance, Stanford's *Named Entity Recognizer* (NER) may be used to identify characters in narratives (Finkel, Grenager, and Manning 2005). However, the Stanford NER only recognizes entities with capitalized names as PERSON. This method may fail to distinguish between narratives that include at least one main character with non-specific labels only (Petersen, Gillam, and Gillam 2008). Moreover, it is not sufficient for certain narratives, such as fairy tales, which may have an animal or a tree as a main character. For this reason, researchers have explored animacy classification as a means of detecting animate objects in narrative (Evans and Orăsan 2000; Karsdorp et al. 2015; Orăsan and Evans 2001). Additionally, others have proposed a case-based reasoning approach to story character recognition, and determined that part-of-speech tagging, WordNet features, and word lists were among the most valuable features for classification (Valls-Vargas, Ontañón, and Zhu 2014). These features, in addition to machine learning, have also been shown to be useful in speaker identification (He, Barbosa, and Kondrak 2013).

Several representations exist for narrative understanding.

The Story Intention Graph (SIG) is an ideally expressive representation which is able to capture implicit information across multiple timelines (Elson 2012). The SIG model is openly accessible, robust across many domains, and designed to reason formally about narrative content. Further, researchers are currently working on automating the SIG encoding process (Elson and McKeown 2010), which lends to their utility in computational narrative complexity assessment.

As of yet, no known system includes modules for analyzing multiple narrative elements to calculate narrative complexity. Here, we present a preliminary approach to this challenge, and evaluate its accuracy using excerpts from fairy tales.

Method

Our system attempts to extract information about and provide a score for twelve of the original thirteen INC narrative elements. The system requires the full text of the story, and a manual annotation for the narrative's SIG encoding. The following sections will explain the processes we use for each of these elements in detail.

Character

A *character* is a reference to the subject of a clause. Following previous work (Evans and Orăsan 2000; Orăsan and Evans 2001), we use part-of-speech tagging and WordNet categories to determine if a particular entity is displaying signs of animacy. For instance, the subject of a clause may engage in dialogue, make a plan using a cognitive verb (e.g., "think"), or take an animated action (e.g., "walk"). If the subject contains a reference that uses capital letters (such as "Tin Soldier" or "Little Tuk"), the system considers the reference a name.

If no character is found, or ambiguous pronouns ("he", "they") are used, the system assigns a score of zero points. One point is awarded if the narrative includes at least one frequent actor with nonspecific labels ("Once, there was a knight."), and two if that character has a name ("Once, there was a knight named Mary."). If more than one named character takes frequent action, three points are awarded.

Setting

The *setting* is any reference to a place or time. Our system identifies setting references by using WordNet's location and time categorization lexnames. To improve the accuracy of the system, state abbreviations that have common duplicate meanings (such as "OH", "OR", "US", and "ME") which WordNet marks as locations were not counted as setting references. Zero, one, and two points are respectively awarded to no, one, or multiple setting references in the text.

Initiating Event

An *initiating event* is a reference to any event or problem that is likely to elicit a character response. Because identifying any sort of problem for a character is a complex narrative understanding problem, our system considers any event in the SIG timeline which prompts a character goal to be an initiating event. We determine whether a goal has been prompted by an event by checking if the two are connected by an *actualizes* link.

Critically, the system must also discover if characters react to the initiating event. These reactions may be *internal responses* or *attempts* (refer to the following sections). If an internal response is marked using WordNet, it is also marked in the SIG encoding during its portion of the timeline. If the internal response directly follows the initiating event in the graph, we consider it to be a reaction. If no initiating events are found, zero points are awarded. One point is awarded if characters do not appear to respond to one or more initiating events. TWo points are awarded if at least one event elicits a response, and three if two or more distinct initiating events elicit a response.

Internal Response

Internal responses encapsulate information about a character's psychological state, such as their emotions and desires. Our system first uses WordNet to determine whether each word expresses (or has a synonym that expresses) a basic emotion. This procedure identifies explicit statements of emotion, such as "The video made him angry". We use the Stanford parser (de Marneffe, MacCartney, and Manning 2006) to determine whether these emotions are related to specific entities. If so, two points are awarded. Otherwise, zero or one points are respectively given to no references or a single reference to emotion.

Plan

A *plan* always includes a cognitive verb that indicates an intent by a character to act on or solve an initiating event. We use WordNet to identify cognitive verbs, as well as several exceptions that indicate intention ("want","hope","desire"). We use the Stanford parser (de Marneffe, MacCartney, and Manning 2006) to determine whether each instance of a cognitive verb is associated with a main character, i.e., the most frequent actor or speaker. Zero, one, two, and three points are respectively awarded to no, one, two, or three cognitive verb references in the text.

Action/Attempt

An *attempt* occurs when a main character acts in response to the initiating event. An *action* may be taken by a main character that is not directly related to the initiating event. Our system examines the SIG encoding for instances of acts with the intent to influence goals associated with the initiating event. These acts must be subsequent to the initiating event itself. If links such as *actualizes ceases*, *prevents*, *attempt to cause*, or *attempt to prevent* exist between the action and the goal, we consider the character to be making an *attempt* with respect to the initiating event. If the character merely acts with no clear connection to the initiating event, we consider these *actions*. If neither actions nor attempts are pursued, zero points are awarded by the system. Instances of only actions taken result in one point awarded. Two points are awarded if at least one attempt is made.

Complication

A *complication* is either (1) an event that prevents the execution of a plan or attempt, or (2) a second initiating event. To determine if the latter is present, the system checks for a second instance of an event that begets a character response (see *Initiating Event*). The system determines if the former occurred by analyzing the SIG encoding for the presence of a *ceases* (or *prevents*) link that is associated with a tagged plan or attempt. No, one, and two points are rewarded for zero, one, and two complications, respectively.

Consequence

A *consequence* is the result of an action or attempt on the initial problem. Such a consequence does not need to resolve the problem, but it must be related to the initiating event and explicitly stated. We can thus make the assumption that consequences arise because the plans of a main character have been achieved or thwarted. To identify such cases, our system looks for Proposition nodes that are connected with initiating event-prompted character goals via *actualizes*, *ceases* (or *prevents*), and *implies*. No, one, two, or three points are awarded for zero, one, two, and three consequences, respectively.

Formulaic Markers

Formulaic markers, such as "once upon a time" or "they lived happily ever after" are used to indicate the beginning or end of a narrative. For each type of these elements, our system assigns zero points for no instance of the element, one point for a single instance, and two points for two or more instances of the element.

Temporal Markers

Temporal markers (e.g., "before", "instantly", "once") indicate time. Like formulaic markers, they may be assessed by simply identifying these key words or phrases. For each type of these elements, our system assigns zero points for no instance of the element, one point for a single instance, and two points for two or more instances of the element.

Causal Adverbial Clauses

Causal adverbial clauses are words or phrases in a sentence which indicate cause (e.g., "because", "since", "so that"). For each type of these elements, our system assigns zero points for no instance of the element, one point for a single instance, and two points for two or more instances of the element.

Knowledge of Dialogue

To determine how many characters are engaging in conversation, we must identify the speakers in the text. In accordance with He et al.'s work (He, Barbosa, and Kondrak 2013), our system extracts speakers by targeting speech verbs ("say", "speak", "talk", "ask", "reply","answer", "add", "continue", "go on", "cry", "sigh", "think") proximal to the utterance. If none of these verbs occur, any verb preceded by a name, personal pronoun, or animate character is chosen. If the story includes named characters, the closest name preceding the pronoun or character label is decided as the speaker. The reference to the speaker is generally extracted by parsing the fragment containing the speech verb, and by following a deterministic method based on syntactic rules. This reference is then matched to known speakers and their associated labels. For example, if the current speaker is "Harry", and "he" is the next speaker, the system will assume "he" refers to "Harry". If no speech verb and character reference is found in the vicinity of the utterance, it is assumed that the quotation marks were used to denote an expression other than dialogue, such as a title. If no dialogue is present, zero points are assigned by the system. If one speaker is present, one point is awarded. If there are two or more characters engaging in conversation, two points are assigned.

Narrative Element	Accuracy
Character	80%
Setting	80%
Initiating Event	75%
Internal Response	55%
Plan	85%
Action/Attempt	85%
Complication	80%
Consequence	90%
Formulaic Markers	95%
Temporal Markers	95%
Causal Adverbial Clauses	100%
Knowledge of Dialogue	95%

Table 1: Evaluation of the automated narrative complexity assessment system.

Evaluation

To evaluate our system, twenty excerpts were randomly selected from open-domain fairy tale texts (Hart 2014). Each selection contained at least 200 characters, and was manually annotated and scored in terms of its index of narrative complexity (Petersen, Gillam, and Gillam 2008). These scores were then compared with the system's assessments (Table 1). The number of points awarded by the system for each narrative element had to agree exactly with the annotation to be considered accurate.

Discussion

We have designed a preliminary, interconnected system that measures elements of narrative complexity. While our system performed moderately well in our preliminary evaluation, there are likely ways to improve its accuracy among all domains. Most significantly, detection of internal response was difficult for our system. Improved narrative understanding of implicit psychological states may increase the accuracy of our system. A machine learning approach may also prove more accurate than relying on WordNet for detecting expressions of emotion.

Another opportunity for improvement might be to replace our means of speaker identification with a supervised machine learning approach, such as that demonstrated by He et

al. (He, Barbosa, and Kondrak 2013), to learn who is speaking. This approach may especially be useful in assessing longer narratives, which may omit speech verbs and rely on reader familiarity of speaker style.

Finally, the present work's attempt at story character identification could also be compared with the case-based approach established by Valls-Vargas et al. (Valls-Vargas, Ontañón, and Zhu 2014), and Karsdorp's animacy classifier (Karsdorp et al. 2015). Further evaluation should determine how well these methods perform with more varied and complex narratives. Future work should also extend this system to evaluate narrator evaluations, and to include automated encoding of story representation to support advanced narrative understanding.

References

Berkman, N. D.; Sheridan, S. L.; Donahue, K. E.; Halpern, D. J.; and Crotty, K. 2011. Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine* 155:97–107.

Birru, M. S.; Monaco, V. M.; Charles, L.; Drew, H.; Njie, V.; Bierria, T.; Detlefsen, E.; and Steinman, R. A. 2004. Internet usage by low-literacy adults seeking health information: An observational analysis. *Journal of Medical Internet Research* 6(3):e25.

Chapman, S. B.; Watkins, R.; Gustafson, C.; Moore, S.; Levin, H. S.; and Kufera, J. A. 1997. Narrative discourse in children with closed head injury, children with language impairment, and typically developing children. *Journal of Medical Internet Research* 6(2):66–76.

de Marneffe, M.-C.; MacCartney, B.; and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Elson, D. K., and McKeown, K. 2010. Building a bank of semantically encoded narratives. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*.

Elson, D. K. 2012. Dramabank: Annotating agency in narrative discourse. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Evans, R., and Orăsan, C. 2000. Improving anaphore resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference*, 154–162.

Fey, M. E.; Catts, H. W.; Proctor-Williams, K.; Tomblin, J. B.; and Zhang, X. 2004. Oral and written story com- position skills of children with language impairment. *Journal of Speech, Language, and Hearing Research* 47(6):1301–1318.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.

Gillam, R. B., and Johnston, J. R. 1992. Spoken and written language relationships in language/learning-impaired and normally achieving school-age children. *Journal of Speech and Hearing Research* 35(6):1303–1315.

Gillam, R. B., and Pearson, N. A. 2004. Test of narrative language. *Topics in Language Disorders*.

Greenhalgh, K. S., and Strong, C. J. 2001. Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools* 32(2):114–125.

Hart, M. 2014. Free ebooks - Project Gutenberg. Gutenberg.org. http://www.gutenberg.org/.

He, H.; Barbosa, D.; and Kondrak, G. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1312–1320.

Karsdorp, F.; van der Meulen, M.; Meder, T.; and van den Bosch, A. 2015. Animacy detection in stories. *Sixth International Workshop on Computational Models of Narrative* 82–97.

Kazantseva, A., and Szpakowicz, S. 2010. Summarizing short stories. *Computational Linguistics* 36(1):71–109.

Koo, M.; Krass, I.; and Aslani, P. 2006. Enhancing patient education about medicines: factors influencing reading and seeking of written medicine information. *Health Expectations* 9(2):174–187.

Mehta, P. D.; Foorman, B. R.; Branum-Martin, L.; and Taylor, W. P. 2005. Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading* 9(2):85–116.

Nation, K.; Clarke, P.; Marshall, C. M.; and Durand, M. 2004. Hidden language impairments in children: Parallels between poor reading comprehension and specific language impairment? *Journal of Speech, Language, and Hearing Research* 47:199–211.

Newman, R. M., and McGregor, K. K. 2006. Teachers and laypersons discern quality differences between narratives produced by children with or without sli. *Journal of Speech, Language, and Hearing Research* 49(5):1022–1036.

Orăsan, C., and Evans, R. 2001. Learning to identify animate references. In *Walter Daelemans and Rémi Zajac, editors, Proceedings of CoNLL-2001*, 129–136.

Pandya, C.; McHugh, M.; and Batalova, J. 2011. LEP Data Brief. Migration Policy Institute, National Center on Immigrant Integration Policy.

Petersen, D. B.; Gillam, S. L.; and Gillam, R. B. 2008. Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders* 28(2):115–130.

Scott, C. M., and Windsor, J. 2000. General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research* 43(2):324–340.

Valls-Vargas, J.; Ontañón, S.; and Zhu, J. 2014. Toward automatic character identification in unannotated narrative text. *Intelligent Narrative Technologies 7,* 38–44.