Categorization of Narrative Semantics for Use in Generative Multidocument Summarization

David K. Elson

Columbia University
450 Computer Science Building
1214 Amsterdam Avenue
New York, NY 10027-7003
delson@cs.columbia.edu

Abstract. The generative summarization of textual stories has been one of the goals of computational narratology since attempts at full semantic NLU in the '70s. Our NLP group has recently created several systems for multidocument news summarization, but using purely statistical methods. Between these poles, there may be an unexplored avenue where knowledge of story structure can give partial, yet useful semantic understanding to a news reader. Such knowledge can then lead to summaries more informed than those based on solely statistical means. This student paper represents work in progress on a two-module system: The first module categorizes news articles into their underlying dramatic structures; the second will attempt to use this understanding to create and execute a generative plan, concisely retelling the story to form a surface-level summary.

1 Introduction

Content selection is a limiting factor in many generation systems, especially those intended for summarization. Statistical summarizers that rely purely on sentence extraction, while practical on unrestricted texts, often produce summaries too "close" to the sources. Without any semantic insight into the gist of the text, a statistical summarizer can favor sentences with nonessential details over those that tell salient events.

While full semantic understanding of an arbitrary text is not a practical goal, the domain of news summarization imposes a restriction on many of its texts that one may be able to exploit: news is, at its core, an act of storytelling. If a news event does not arouse fear, or compassion, or a laugh, or some other emotional response to the dramatic underpinnings of the story, it does not make the pages of a broadsheet.

In this paper, we describe the initial stages of a project to create a new model of the rhetorical structure of textual stories and apply it to the content selection piece of a generative multidocument news summarizer.

A. Belz et al. (Eds.): INLG 2004, LNAI 3123, pp. 192–197, 2004.

2 Related Work

The link between news and storytelling has been investigated before, originally by those attempting full semantic understanding in the '70s [17]. More recently, researchers such as Power [13], Mann and Thompson [9] have developed rhetorical structure theory as a more practical, high-level representation of a general text's structure. Plans have been used by Hovy [7], Moore and Paris [12] to improve text generation for expert systems, as have schemas by McKeown [11].

In parallel with this work, computational narratology has aimed to capture the essence of drama for both story understanding and generation. The different approaches vary widely: Some place the notion of conflict between characters at the atomic center of their models [18]; others prefer surface-level syntactic structure [16], the autonomy of self-determining characters [15], the human process of storytelling [8] or even the reader's emotional response to a story [2].

Computational narratology has experienced a revival in recent years, by both theorists and those bringing practical applications to bear. The introductory paper by Mateas and Sengers [10] from the 1999 AAAI Fall Symposium gives an excellent overview, as does the summary of the MIT Media Lab's reading group on narrative intelligence by Davis and Travers [4]. The potential for a fusion with rhetorical structure theory for summarization has been little explored.

3 System Design

Telling interesting narratives is a skill that journalists and writers of fiction alike are trained to do. They each look for the best "angle" with which to cast the events of their worlds into a dramatic mold.

One of the motifs in the heritage of literary theory, where computational limitations are not even considered, has been the idea that there are only so many stories that *can* be told ([1], [14]). Within the news domain, there are even fewer, as journalists strive to sell their stories to the public by invoking plots, characters, and themes that have been shown to generate interest in the past (e.g., the *death of the unsuspecting innocent*; the *fall from power*; the *personal attack*; the *rich magnate*; the *comeback* or *recovery*; *justice for the evil*; *new technology for old wants*).

Each of these plots, characters and themes are metadata for a story that carry certain narrative connotations. For example, a *fall from power* article probably describes the person in question, portrays the person as a villain brought to justice or a hero wrongfully hindered, exposits the reason for the fall, and characterizes the opposing force. As these are the essentials of the *fall from power* story, they represent the facts most important to the reader, and as such, they are the best facts to portray in a summary. If, by contrast, an article tells of the *death of the unsuspecting innocent*, the corresponding content selection problem is slightly different.