Understanding and summarization

Richard Alterman
Computer Science Department, Brandeis University, USA

Abstract. This article is an overview of the literature on narrative summarization. The capacity to summarize is a fundamental property of intelligence and has significance for several areas of artificial intelligence research and development. The first part of the paper includes a description of four critical features of a summary. The bulk of this review is concerned with sorting available summarization frameworks and techniques. A latter section of the paper describes the significance of summarization technology to three current topics in artificial intelligence: explanation-based learning, case-based reasoning, and plan evaluation.

1 Introduction

This paper is a review of the literature on summarization. The basic idea of summarization is to take a body of information and reduce its size and content to its important points. The capacity to summarize is a fundamental property of intelligence. In our daily lives there is an overwhelming amount of information to process and much of it is neither relevant nor of interest. Summarization processes allow an intelligent agent to focus on the most significant aspects of a given understanding.

A pragmatic reason for studying summarization is that it provides a useful way to report back on a body of knowledge. Take a simple example: suppose a user sits down to an information retrieval system and requests all information relevant to Japanese embargos of imports. In a given knowledge-base there might be an overwhelming number of episodes to describe. What the user does not want is a detailed description of each episode. Rather, s/he would prefer a summary of the more important episodes.

Summarization also serves an important function in research on cognition. One of the central issues of cognitive science is characterization of the 'understanding' process. Summarization provides a test of a given model of understanding. Suppose a researcher claims that the thematic level of understanding plays a significant role in understanding. Summarization provides a test and a methodology for exploring this sort of claim. If thematic understandings are important, then they will be reflected in the sorts of summaries that human subjects produce for a given text and thus should also be reflected in summaries produced by computer models.

Summarization is not a single phenomenon. There are many different kinds of...
summaries. To name just a few, there are abstracts, epitomes, overviews, abridgments, digests, and recapitulations. Each style of summarization requires a slightly different viewpoint on extracting the essential content of a given text or understanding.

Different approaches to summarization emphasize alternate effects and functions of the process. Summarization is sometimes treated as a problem of memory (e.g. Kintsch & van Dijk, 1978) i.e., what does the subject remember of the text after various periods of time. The point here is that as, for example, a story is read it is not perfectly stored in memory, but only its most significant parts are retrievable; by what process does this occur? Other models of summarization are biased towards one or other implicit structure of the text. Structures that have been previously computationally tested for their contribution to the summarization process include: story schema (Rumelhart, 1975), schema narrative trees (Simmons & Correia, 1980; Correia, 1980), sketchy scripts (DeJong, 1979), plot units (Lehnert & Loiselle, 1989; Lehnert, 1981), story points (Wilensky, 1980), and generic knowledge structures (Graesser & Clark, 1985). Each of these structures presents a different point of view on the underlying understanding and summarization processes.

A first pass approximation of some important properties of a summary, includes the following:

• does the summary reduce the workload for the interpreter/understander over the text?
• does the summary maintain coherence?
• does the summary maintain coverage?
• does the summary include the important events of the story?

The issue of workload suggests that it should take less work to construct an interpretation of a summary than the original text. (A summary is not only shorter but it is also, in some manner, ‘simpler’.) The question of coherence suggests it is not good enough to just reduce the quantity of text — the summary must hold together and make sense. The third question (coverage) indicates that a good summary must cover, at least implicitly, many of the events of the original text. The last issue (importance) indicates that the summary should include the important parts of the text, and, where they are not necessary for reasons of coherence, exclude the unimportant parts.

The remainder of the article begins by expatiating on the axis role of representation in computational models of understanding and summarization. The bulk of the paper will be concerned with sorting the available summarization frameworks and techniques. Although traditionally summarization has been studied in the context of narratives, a latter section of this paper will describe why summarization research is significant for three other areas of artificial intelligence: explanation based learning, plan evaluation, and case-based reasoning.

2 Representation. The artefact of ‘understanding’

What does it mean to say that a machine has modelled ‘understanding’? The traditional answer in artificial intelligence (AI) models of text comprehension is to