Exploiting Captions for Multimedia Data Mining

Neil C. Rowe
U.S. Naval Postgraduate School, USA

INTRODUCTION

Captions are text that describes some other information; they are especially useful for describing non-text media objects (images, audio, video, and software). Captions are valuable metadata for managing multimedia, since they help users better understand and remember (McAninch, Austin, & Derks, 1992-1993) and permit better indexing of media. Captions are essential for effective data mining of multimedia data, since only a small amount of text in typical documents with multimedia—1.2% in a survey of random World Wide Web pages (Rowe, 2002)—describes the media objects. Thus, standard Web browsers do poorly at finding media without knowledge of captions. Multimedia information is increasingly common in documents, as computer technology improves in speed and ability to handle it, and as people need multimedia for a variety of purposes like illustrating educational materials and preparing news stories.

Captions also are valuable, because non-text media rarely specify internally the creator, date, or spatial and temporal context, and cannot convey linguistic features like negation, tense, and indirect reference. Furthermore, experiments with users of multimedia retrieval systems show a wide range of needs (Sutcliffe et al., 1997) but a focus on media meaning rather than appearance (Armitage & Enser, 1997). This suggests that content analysis of media is unnecessary for many retrieval situations, which is fortunate, because it is often considerably slower and more unreliable than caption analysis. But using captions requires finding them and understanding them. Many captions are not clearly identified, and the mapping from captions to media objects is rarely easy. Nonetheless, the restricted semantics of media and captions can be exploited.

FINDING, RATING, AND INDEXING CAPTIONS

Background

Much text in a document near a media object is unrelated to that object, and text explicitly associated with an object often may not describe it (i.e., "JPEG picture here" or "Photo39573"). Thus, we need clues to distinguish and rate a variety of caption possibilities and words within them, allowing for more than one caption for an object or more than one object for a caption. Free commercial media search engines (i.e., images.google.com, multimedia.lycos.com, and www.altavista.com/image) use a few simple clues to index media, but their accuracy is significantly lower than that for indexing text. For instance, Rowe (2005) reported that none of five major image search engines could find pictures for "President greeting dignitaries" in 18 tries. So research is exploring a broader range of caption clues and types (Mukherjea & Cho, 1999; Sclaroff et al., 1999).

Sources of Captions

Some captions are explicitly attached to media objects by adding them to a digital library or database. On Web pages, HTML "alt" and "caption" tags also explicitly associate text with media objects. Clickable text links to media files are another good source of captions, since the text must explain the link. A short caption can be the name of the media file itself (e.g., "socket_wrench.gif").

Less explicit captions use conventions like centering or font changes to text. Titles and headings preceding a media object also can serve as captions, as they generalize over a block of information, but

they can be overly general. Paragraphs above, below, or next to media also can be captions, especially short paragraphs.

Other captions are embedded directly into the media, like characters drawn on an image (Lienhart & Wernicke, 2002) or explanatory words at the beginning of audio. These require specialized processing like optical character recognition to extract. Captions can be attached through a separate channel of video or audio, as with the "closed captions" associated with television broadcasts that aid hearing-impaired viewers and students learning languages. "Annotations" can function like captions, although they tend to emphasize analysis or background knowledge.

Cues for Rating Captions

A caption candidate's type affects its likelihood, but many other clues help rate it and its words (Rowe, 2005):

- Certain words are typical of captions, like those having to do with communication, representation, and showing. Words about space and time (e.g., "west." "event," "above," "yesterday") are good clues, too. Negative clues like "bytes" and "page" can be equally valuable as indicators of text unlikely to be captions. Words can be made to be more powerful clues by enforcing a limited or controlled vocabulary for describing media, like what librarians use in cataloging books (Arms, 1999), but this requires cooperation from caption writers and is often impossible.
- Position in the caption candidate matters: Words early in the text are four times more likely to describe a media object (Rowe, 2002).
- Distinctive phrases often signal captions (e.g., "the X above," "you can hear X," "X then Y") where X and Y describe depictable objects.
- Full parsing of caption candidates (Elworthy et al., 2001; Srihari & Zhang, 1999) can extract more detailed information about them, but it is time-consuming and prone to errors.
- Candidate length is a clue, since true captions average 200 characters with few under 20 or over 1,000.
- A good clue is words in common between the candidate caption and the name of the media file, such as "Front view of woodchuck burrowing" and image file "northern_woodchuck.gif."
- Nearness of the caption candidate to its media actually is not a clue (Rowe, 2002), since much nearby text in documents is unrelated.
- Some words in the name of a media file affect captionability (e.g., "view" and "clip" as positive clues and "icon" and "button" as negative clues).
- "Decorative" media objects occurring more than once on a page or three times on a site are 99% certain not to have captions (Rowe, 2002). Text generally captions only one media object except for headings and titles.
- Media-related clues are the size of the object (small objects are less likely to have captions) and the file format (e.g., JPEG images are more likely to have captions). Other clues are the number of colors and the ratio of width to length for an image.
- Consistency with the style of known captions on the same page or at the same site is also a clue because many organizations specify a consistent "look and feel" for their captions.

Quantifying Clues

Clue strength is the conditional probability of a caption given appearance of the clue, estimated from statistics by c/(c+n), where c is the number of occurrences of the clue in a caption and n is the number of occurrences of the clue in a noncaption. If we have a representative sample, clue appearances can be modeled as a binomial process with expected standard deviation $\sqrt{cn/(c+n)}$. This can be used to judge whether a clue is statistically significant, and it rules out many potential word clues. Recall-precision analysis also can compare clues; Rowe (2002) showed that text-word clues were the most valuable in identifying captions, followed in order by caption type, image format, words in common between the text and the image filename, image size, use of digits in the image file name, and image-filename word clues.

Methods of data mining (Witten & Frank, 2000) can combine clues to get an overall likelihood that