Languages & the Media Conference, Berlin, October 2018

From 'Slicing bananas' to 'Pluto the Dog': human and automatic approaches to visual storytelling

Sabine Braun, Kimm Starr, University of Surrey

Recently there has been a surge of interest in methods for describing audiovisual content, be it for the purposes of automatic image search and retrieval, to advance visual storytelling, or through increasing demand in described content following changes in national and European broadcasting legislation to meet the needs of visually impaired audiences. Although the computer vision and natural language processing communities have intensified research into the automatic generation of descriptions (Bernardini et al. 2016), even the automatic description of still images remains challenging in terms of accuracy, completeness and robustness (Husain & Bober 2016), whilst descriptions of moving images and visual storytelling pose additional challenges linked to temporality, including co-referencing (Rohrbach et al. 2017) and other features of narrative continuity (Huang et al. 2016). Despite rapid advances in machine learning, automatically generated descriptions are currently at best plainer than their human equivalents, which tend to be grammatically and stylistically more complex; but often the automatic versions are also incorrect and incoherent such that, depending on available training data, any oblong yellow object (including Disney's Pluto) may be identified as a banana, and that actions or links between scenes in moving images may be absent. By contrast, human-made audio descriptions, originally aimed at visually impaired audiences, provide one of the most elaborate and reliable types of content description currently available for (still and) moving images, but are expensive to produce. However, together with other material such as programme guides, subtitles and film scripts, human audio descriptions provide a rich source of information about visual, auditory and verbal elements in audiovisual content that can be exploited for research and machine training.

Against this backdrop, this presentation will report on a study that is currently conducting a systematic comparison of human descriptions of audiovisual content with corresponding machine-generated descriptions, with the aim of identifying key characteristics and patterns of manually and automatically produced descriptions, and evaluating each method. The focus of the presentation will be on the preliminary outcomes of the comparison, drawing on corpus-based and discourse-based approaches to analyse, for example, how each method handles character identification, (recurrent) references to simple and complex objects, focalisation and other elements that are crucial to achieving narrative continuity for visual storytelling. This will be grounded in discourse-based models of human comprehension and processing of narrative audiovisual content. By way of a preliminary evaluation of the two methods (i.e. human and machine-generated description), we will discuss how human understanding of narrative audiovisual content and of content enrichment needs, and human techniques and strategies of description can inform and guide the development of models of automatic description.

The broader aim of this work is to advance current understanding of multimodal content description and to contribute to enhancing and personalising content description services and technologies. This will benefit the Creative Industries, especially TV broadcasters and on-demand media service providers, as well as people using their services.

This study is part of the European H2020 project MeMAD (Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy; https://memad.eu*) and is financially supported by the European Commission.*