Using Rhetorical Annotation in TrecVID BBC Rush Summarization Task

- A BLUEBOOK NOTE -

Zeljko Obrenovic

Created: **11/09/2007**, Last modified: **12/09/2007**

As one of the task, TrecVID runs BBC rushes summarization. In this blue book note, we describe some ideas how rhetorical annotation can be used in this task. In brief, the idea is to reuse Stefano's work from VoxPopuli, and to try to apply it to the generation of video summarizations for BBC rushes. VoxPopuli basically solves similar problem: generating (biased) short videos from unedited video material.

Our potential participation in this task is interesting for several reasons:

- *We have access to 100 hours of BBC series*
- *Other partners will generate some (mostly low-level) annotation which we can reuse*
- *This could also be interesting application if my VENI proposal is accepted*

In this blue book note we describe the background, and discuss three groups of ideas about how we can be involved with this task: rushes summarization as media production process: Relation with canonical processes of media production; rlation with Stefano's Work, and relations with NewsML.

Background

See: http://www-nlpir.nist.gov/projects/tv2007/tv2007.html

Rushes are the raw material (extra video, B-rolls footage) used to produce a video. 20 to 40 times as much material may be shot as actually becomes part of the finished product. Rushes **usually have only natural sound**. **Actors are only sometimes present**. So **very little if any information is encoded in speech**. Rushes contain many frames or sequences of frames that are **highly repetitive**, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies over, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as **stock footage - reusable shots of people, objects, events, locations**, etc. Rushes may share some characteristics with "ground reconnaissance" video.

The BBC Archive has provided about **100 hours of unedited material in MPEG-1 from about five dramatic series**. Most of the videos have durations of about 30 minutes. Half the videos will be used for systems development and half reserved for system test.

Sample ground truth - lists of important segments identified by major objects/events - will be created by Dublin City University for some development clips and provided with the development data. These will be just examples and not intended as training data. The ground truth for the test data, created by the same process/people will be the basis for the evaluation.

The system task in rushes summarization will be, given a video from the rushes test collection, to **automatically create an MPEG-1 summary clip less than or equal to 4% of the original video's duration**. This means the average summary will be less than or equal to 60 seconds long. The summary should show the main objects (animate and inanimate) and events in the rushes video to be summarized. The summary should minimize the number of frames used and present the information in ways that maximizes the usability of the summary and speed of objects/event recognition.

Such a summary could be returned with each video found by a video search engine much as text search engines return short lists of keywords (in context) for each document found - to help the searcher (whether professional of recreational) decide whether to explore a given item further without viewing the whole item. It might be input to a larger system for filtering, exploring and managing rushes data.

Although in this pilot task we limit the notion of visual summary to a single clip that will be evaluated using simple play and pause controls, there is **still room for creativity in generating the summary**. Summaries **need NOT be series of frames taken directly from the video to be summarized and presented in the same order**. Summaries can contain *picture-in-picture*, *split screens*, and results of other techniques for organizing the summary. Such approaches will raise **interesting questions of usability**.

Main Issues

We have three groups of ideas about how we can be involved with this task:

- Rushes summarization as media production process: Relation with canonical processes of media production
- Relation with Stefano's Work,
- Relations with NewsML.

Rushes summarization as media production process: Relation with canonical processes of media production

The idea here is to describe the rushes summarization in terms of canonical processes of media production. Our contribution could be providing such a description at the early stage so that partners can get the overview of the system before they get too involved in the implementation details.

Our hypothesis is that if we use canonical processes to describe this task at early stage, we could guide the design of resulting system by helping the developers to fit their work into bigger context, collaborate with each other, integrate with existing systems, and generate new ideas. The timing is not perfect, but this could also be an excellent test case for our main canonical process paper, illustrating the benefits of using canonical processes.

Some ideas how this mapping can be done. We should probably start by defining a generic model of rush summarization, which could then be specialized by partners to instantiate different forms of rush summarization.

Ideas for initial mapping to canonical processes:

- **Premeditate**, definition of the summarization task, identification of available content and other resources. Here we see what we get and put it on the table to get the ideas. This is basically the input for construct message process.
- **Create**, many new media items will be created by transforming or merging frames selected from the rushes. Items can also be transformed to increase the contrast or change the contrast to make it more accessible.
- **Annotate**, we inherit some annotation from BBC, low-level annotation will be created related to detection of scene boundaries and identification of objects in the scenes. High-level rhetorical is desirable, and probably has to be added manually.
- **Package**, selected frames has to be grouped according to the defined summarization task. Grouping can also involve addition of other outside content, such as NewsML.
- **Query**, several querying interfaces are needed. The basic one is querying for initial content of frames and BBC annotation. After our annotation is added, we also need the interface for that. It probably cannot be the same interface.
- **Construct message**, the concrete systems will different according to which message they want to convey. We would like to work on something similar to Stefano's work, i.e. creating biased presentation. The concrete bias is our message. Other partners may be interesting in other aspects, such as novelty, visual harmony (show all red scenes, for example), or many other things…
- **Organise**, we probably need some generic description of presentation that can be shared among partners. Organize should define generic presentation, but also enable combining the presentation from different partners in one coherent presentation. For that we need standard exchange format, and merge operators.
- Publish, and
- Distribute.

Relation with Stefano's Work

Main idea is to try to apply Stefano's approach, used in VoxPopuli to help the summarization. VoxPopuli basically solves similar problem: generates interview videos from unedited video material. However, there are lots of differences, and the Stefano's approach cannot be directly applied. Here are some questions and problems:

- **What kind of rhetorical annotation is appropriate for BBC rushes**? Using the Toulmin model of argumentation, which is the basis for Stafano's work, is probably not the most appropriate for rushes

summarization. Stefano's approach cannot be directly applied because VoxPopuli uses encoded verbal information contained in the audio channel. Rushes, however, contain almost no speech. Identifying the claims and the argumentation structures in rushes may be hard or practical impossible.

- **What is the desired effect, i.e. what is the goal of summaries?** In other words, how the summaries can be biased. The official task does not limit or bias this, most ideas are related about identification of content.
- Using rushes as stock footage - reusable shots of people, objects, events, locations, looks very interesting. **Is it possible to describe de story in generic terms** (people, objects, events, locations) and automatically instantiate this generic template with instance from rushes? **Can existing documentaries, such as that from Stafno, be enriched with such footage**, e.g. when the interviewer mention the people on the street, we can show one such a scene. What kind of annotation is necessary on both sides if such merge is to be achieved?
- **Usability and accessibility issues…**

Relations with NewsML

Two topics:

- Finding and associating news related to the object within the rushes (about actors, locations, the episode itself)
- Creating presentation where nes can be illustrated with some content from rushes, i.e. o using rushes as stock footage.

More after K-Space meeting in Nice…

References

- TrecVID 2007, http://www-nlpir.nist.gov/projects/tv2007/tv2007.html
- VOX POPULI Demo Page, http://homepages.cwi.nl/~media/demo/IWA/
- The Toulmin Model of Argumention
  http://en.wikipedia.org/wiki/Stephen_Toulmin#The_Toulmin_Model_of_Argument
  http://commfaculty.fullerton.edu/rgass/toulmin2.htm