Vox Populi: a Tool for Automatically Generating Video Documentaries

Stefano Bocconi
CWI
P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands
First.Last@cwi.nl

Frank Nack
CWI
P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands
First.Last@cwi.nl

Lynda Hardman*
CWI
P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands
First.Last@cwi.nl

ABSTRACT

Vox Populi is a system that automatically generates video documentaries. Our application domain is video interviews about controversial topics. Via a Web interface the user selects one of the possible topics and a point of view she would like the generated sequence to present, and the engine selects and assembles video material from the repository to satisfy the user request.

Categories and Subject Descriptors

H.5.4 [**Hypertext,Hypermedia**]: Architectures, Navigation, User issues; I.7.2 [**Document Preparation**]: Hypertext/hypermedia, Multi/mixed media

General Terms

Design, Experimentation, Human Factors

Keywords

Hypermedia, Automatic Linking, Structured Annotations, Thesaurus

1. OPINIONATED VIDEO SEQUENCES

Vox Populi is a system we are using to explore the field of automatic generation of video documentaries as a new way for documentarists to make their material available to viewers. Figure 1 presents a screenshot of the interface[1]. The video material is provided by a group of independent amateur filmmakers who created the documentary "Interview With America" (IWA)[2]. The 8 hours of material in the IWA database contains interviews with United States residents from different socio-economic groups on the events happening after the terrorist attack on the 11th of September 2001. Issues discussed include the war in Afghanistan, anthrax, media coverage and social integration in multicultural societies.

Vox Populi seeks to avoid the difficulties encountered while trying to create a final version of a documentary. Different opinions of the editors inevitably clash when a choice has to be made about what material should be selected and what should be left out. An alternative solution is to create a mechanism able to show all material to the viewer in different documentaries, generated according to different topics of interest to the user. This approach shifts the burden of the documentarist from making all the editorial choices for a single version to annotate the material to allow multiple versions. To reflect the highly controversial nature of the subject, the generated video sequence creates the impression of an argumentation dialogue between different interviewees, which is a form of narrative generally more appealing to a viewer than a sequence of persons speaking about a common topic.

To illustrate the current system we present an example scenario. The user starts by selecting an opinion (the left menu box in Figure 1a), which is a topic (e.g. the War in Afghanistan) and a position with respect to the topic (pro, against or neutral). The user is then presented with a list of interviewees (the right menu box in the figure) who are expressing the selected opinion. The description in the menu is meant to give an idea about the interviewee. After having selected a particular interviewee (e.g. "lawyer in Harward"), the user is presented with another choice: whether the selected opinion must be supported or contradicted by the generated video documentary. Let us suppose that the user chooses to contradict the lawyer's position. The original statement of the lawyer is: "I am not a fan of military action, but in the current situation I cannot think of a more effective solution". To contrast her point of view, the engine chooses a shop owner saying "War has never solved anything" (Figure 2, segment 2) and an Empire State Building employee saying "They are using two billion dollar bombs on ten dollar tents" (segment 4). The final generated video sequence becomes: the lawyer saying "I am not a fan of military action" (segment 1); the shop owner saying "war has never solved anything" (segment 2); the lawyer saying "in the current situation I cannot think of a more effective solution" (segment 3); the Empire State Building employee saying "two billion dollar bombs on tents" (segment 4).

The output of the process is a SMIL file that can be played by a SMIL player such as Real Player (as used for the screenshots in Figure 2).

In the following we will describe step by step how a video sequence is generated, explaining briefly the underlying tech-

*Lynda Hardman is also affiliated with the Technical University of Eindhoven.

[1]The demo is accessible online from the demo page at `http://www.cwi.nl/~media/demo/VoxPopuli/`

[2]`http://www.interviewwithamerica.com/documentary.html`

HT'05, September 6–9, 2005, Salzburg, Austria.
ACM 1-59593-168-6/05/0009.

niques: the annotation schema in Section 2, the selection of video segments in Section 3 and the editing based on film theory/narrative in Section 4. Further references can be found (together with more information about the technique used), in [1].

2. ENCODING VIDEO SEMANTICS

Vox Populi assembles a video sequence in the following steps: first an interview is selected by the user, then other video segments are selected according to the goal (Section 3) and finally the sequence is assembled (Section 4).

To enable the selection process, Vox Populi must "understand" the semantics contained in the video material. We defined rhetoric annotations to encode the verbal information contained in the audio channel. Since annotations are very time-consuming, our own annotation schema is designed to be not too complex for annotators. The main elements of this schema are the statements and the rhetoric structure in which the statements are placed, i.e. the Toulmin Model. A statement is intended to capture the semantics of a claim an interviewee makes and it is composed of a `subject`, a `modifier` and a `predicate`. The `subject(s)` represents the subject of the statement, the `predicate(p)` qualifies the `subject` and the `modifier(m)` modifies the relation between the `subject` and the `predicate`. The statement "Two billion dollar bombs on tents", for example, is encoded as `s:Bombing m:not p:effective`. Each statement is associated with at least one video segment, and each video segment is annotated with at least one statement.

Each term used as a value in one of the three parts of a statement belongs to a thesaurus. We use a thesaurus relating the terms with the canonical relations *synonym* (or *similar*), *antonym* (or *opposite*), *hypernym* (or *generalization*) and *hyponym* (or *specialization*). The rationale behind the use of a thesaurus is that the relation between two terms can be used to infer the relation between two statements that contain those terms. For example, if `Bombing` has relation *opposite* to `Economic Aid`, then "Bombing effective" is contraddicting "Economic Aid effective".

Usually an interviewee uses different statements to state and support her opinion. We use the Toulmin Model to encode the role these different statements play in building the full claim. Using this model, an argument is broken down into its functional components: the claim made, the grounds supporting it (i.e., facts to support the claim), a warrant for connecting the grounds to the claim, a backing (the theoretical or experimental foundations for the warrant), qualifiers (some, many, most, etc.) that strengthen or weaken the claim, and rebuttals, such as concession (contradicts but is less strong than the claim) or condition (that, if true, could invalidate the claim).

For the current discussion it is important to notice that a statement with role claim, data, warrant or backing contributes positively to the general claim made by the interviewee, while a statement with role concession or condition weakens it (like a rebuttal). This is used by Vox Populi when determining how to support or contradict the user-chosen interview, as explained in the following section.

3. SELECTING VIDEO SEGMENTS

Using the encoding of the statement and the relations between terms in the thesaurus, the video segments in the repository are related to each other. For two related video segments A and B one of the following is true: either A supports B or A contradicts B.

Vox Populi uses this information to compose short video documentaries. The interview selected by the user is decomposed in its video segments and related statements, and for each video segment the system juxtaposes other video segments based on the following: if the Toulmin role is condition or concession, retrieve supporting video segments if the strategy is contradict or retrieve contradicting video segments if the strategy is support. If the Toulmin role is claim, data, warrant or backing, retrieve contradicting video segments if the strategy is contradict, or retrieve supporting video segments if the strategy is support.

Referring to the example introduced in Section 1, Vox Populi selects the video segment of the shop owner because it supports the lawyer's statement "I am not a fan of military action" (concession in Toulmin model) and the video segment of the Empire State Building employee because it contradicts the lawyer's statement "in the current situation I cannot think of a more effective solution" (the interviewee's claim).

In [1] we describe how the generation mechanism can be influenced by the author/annotator.

4. ADVANCED EDITING

The selection presented in Section 3 is based on what in rhetoric is called the logos, i.e. the logic and the meaning of the words the interviewees use during the interviews. Human-edited documentaries are not only based on this principle, and more disciplines come into play when editing video material, such as film and narrative theory. We are creating rules that implement principles borrowed from those disciplines. However, the various rule sets might clash: for example the rhetoric rule set might require that two video segments (one a close-up and another a long shot) should directly follow each other, but the film theory rule set could require shot continuity and not allow a juxtaposition of differently framed segments.

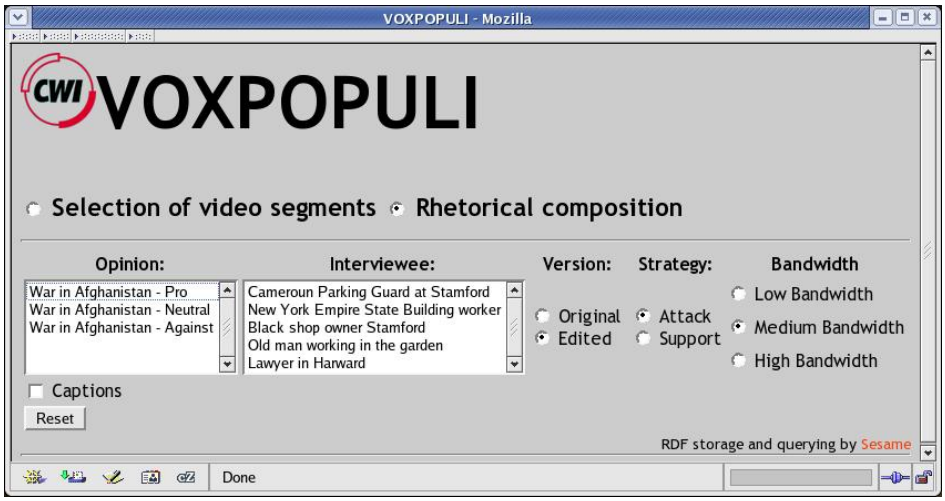
To solve this problem we have implemented an architecture where rules must be sequenced by the author in order of importance. Vox Populi applies them, evaluating the result based on how many rules it could fulfill. Rules higher up in the chain have more weight than rules further down the chain and the author can set a threshold to indicate whether the engine should try hard to apply all the rules or stop when some of them are satisfied. In this way the documentarist can choose an editing style emphasizing either rhetoric or film theory, by ordering the rules accordingly, and define by using the threshold how good or, since quality requires more computation time, how fast the result must be.

Acknowledgments

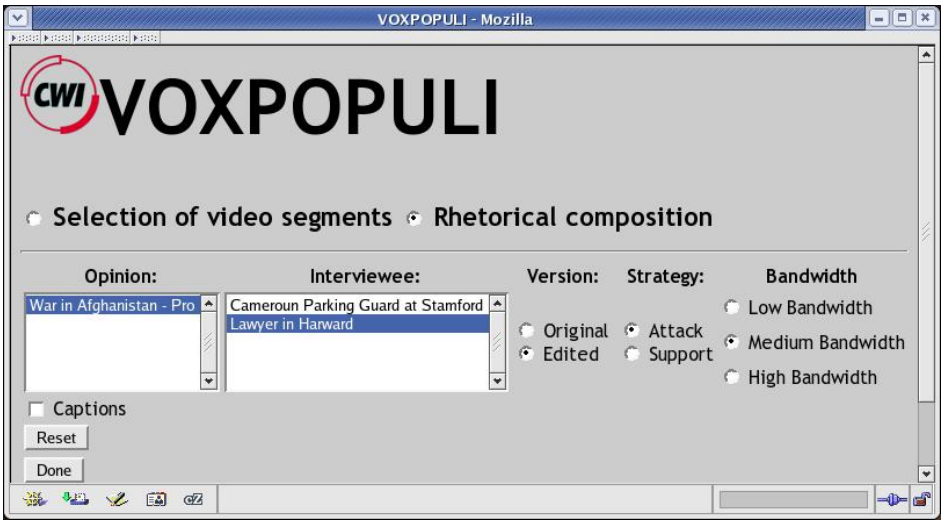
This research was funded by the Dutch national NWO ToKeN I^{2}RP and CHIME projects.

5. REFERENCES

[1] S. Bocconi, F. Nack, and L. Hardman. Supporting the Generation of Argument Structure within Video Sequences. In *Proceedings of the sixteenth ACM Conference on Hypertext and Hypermedia 2005*, September 2005. In press.



(a) The initial page



(b) After user selection

Figure 1: Vox Populi Web Interface

1. Lawyer: *I am not a fan of military action*
2. Shop owner: *war has never solved anything*
3. Lawyer: *in the current situation I cannot think of a more effective solution*
4. Empire State Building employee : *two billion dollar bombs on tents*

Figure 2: A generated sequence, numbered left-to-right