USING RHETORICAL ANNOTATIONS FOR GENERATING VIDEO DOCUMENTARIES

Stefano Bocconi, Frank Nack, Lynda Hardman

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT

We use rhetorical annotations to specify a generation process that can assemble meaningful video sequences with a communicative goal and an argumentative progression. Our annotation schema encodes the verbal information contained in the audio channel, identifying the claims the interviewees make and the argumentation structures they use to make those claims. Based on this schema, we construct a semantic graph which is traversed by rhetoric-based strategies selecting video segments. The selected video segments are edited to form a meaningful video sequence.

1. INTRODUCTION

Our goal is to generate video sequences that have a specific communicative goal and an argumentative progression, just as human-created video documentaries. This requires a system to understand that there are different views on a subject and that points can be made. We present an annotation schema that makes it possible for our generation engine, Vox Populi, to generate argumentative video sequences.

In this way a documentarist can focus on the collection of a rich information set and its annotations and let the system automatically assemble video sequences for the audience to explore the material. This becomes a more effective way of conveying information only, however, if the presented result facilitates presentations beyond a simple sequence of potentially related interviews, which is less appealing to the user.

Vox Populi utilizes an audio-visual repository to automatically generate short video documentaries that make a point and show argumentative progression. Though we are interested in the visual material as well, for the generation of a video argument we focus in this paper on the verbal information contained in the audio track.

The video material is provided by Interview With America (IWA)[1], a documentary shot by a group of independent amateur filmmakers. 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.

The structure of the paper is as follows. Section 2 provides a scenario to introduce the Vox Populi environment. We explain the theory on which our rhetoric engine is based in section 3. We then detail the annotation structures we use to manually describe the claims made in interviews (section 4) and how automatic editing can exploit these structures for the generation of novel lines of argument (section 5). Conclusions are given in section 6 and section 7 outlines future work.

2. SCENARIO

To facilitate the presentation, we introduce the functionality our engine[2] can provide with a simplified example. Imagine a user willing to view arguments from the IWA repository that feature at least one interviewee in favor of the war in Afghanistan. The engine will first establish the key interview and then start constructing an argument around it. The way our engine does that is, for example, by selecting content that expresses a contrasting point of view. In Figure 1, for example, the original selected interview features a young woman (on the right) saying: ”I am not a fan of military actions, but in the current situation I can not think of a more effective solution”. To express a contrasting point, the engine selects other interview segments (on the left of the figure) and edits a sequence which is visually represented by the lower part of Figure 2 and verbally by the following: ”I am not a fan of military actions - war has never solved anything - in the current situation I can not think of a more effective solution - two billions dollars on tents”.

3. THE MODEL FOR ARGUMENT STRUCTURE

The argument structure generated by Vox Populi is based on claims. A claim can be in the form of a single statement or be part of a larger structure in which additional statements

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

[2] Vox Populi is implemented in Java and its output is encoded in SMIL2. A demo can be found at `http://www.cwi.nl/~media/demo/VoxPopuli/`

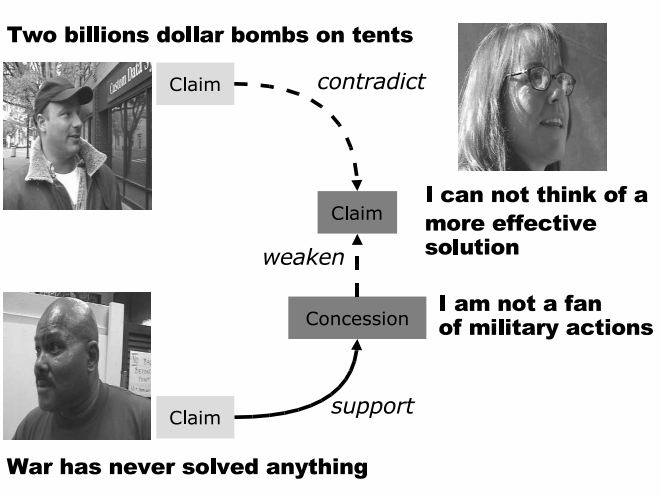

Fig. 1. An example from the IWA repository

support the claim. We use the Toulmin Model [1] because it describes the general structure of rational argumentation. In 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, like concession (contradicts but is less strong than the claim) or condition (that, if true, could invalidate the claim).

The Toulmin model identifies the different discourse parts used to make a claim and their role. In section 2 we showed how Vox Populi uses this knowledge: to find a statement that can contradict the opinion expressed by the woman, the system retrieves a statement that supports the concession, since the latter weakens the claim according to the Toulmin model. Other argumentation systems like [2] use Toulmin in a similar way.

4. ANNOTATIONS - IDENTIFYING ARGUMENT UNITS

We distinguish two types of annotations: descriptive and rhetorical. Our descriptive annotations cover the who, where, when, and what in the video and are in line with those suggested by [3] and [4]. An example query which requires such annotations could be "Select all the answers to the question X given by people of race Y and level of education Z". For our task they are not sufficient, though, because they only support clustering of the sequences, but not building of argumentation structures.

We claim that annotations that capture the rhetoric semantics of a statement facilitate the generation of video sequences with an argumentative progression. Figure 1 provides an example how a dispute can be realized. We have the following requirements for our annotation schema:

1. not too cumbersome for the annotator, because annotations are mostly manually made. The annotator is in our case the documentarist or the person that wishes to make the video material available.
2. expressive enough to capture the semantic of natural language claims (the rhetoric intention of the statements
3. defined formally to be used by the inferencing mechanism described in 5.1

In the following we introduce the novel annotation structures we use.

4.1. Statements

The smallest unity that we annotate is the **video segment**, i.e. that part of the video footage where an interviewee makes a **statement**, such as the following: "I am never a fan of military actions, in the big picture I do not think they are ever a good thing". A single video segment can be shown to a viewer who would understand the meaning expressed in it.

To describe statements, we found that a triplet structure, composed by `subject`, `modifier` and `predicate`, represents a good trade-off solution for the requirements we set in section 4. The subject (`s`) represents the subject of the statement, the predicate (`p`) qualifies the subject and the modifier (`m`) values the relation of the subject with the predicate. The above-mentioned statement is, for example, encoded as `Military Action(s) never(m) effective(p)`.

4.2. The vocabulary

In each statement `s`, `m` and `p` are instantiated with terms from a domain-dependent vocabulary chosen by the annotator. An annotator can build the vocabulary while annotating the material with the terms she uses to compose the statements, or she can use an existing taxonomy, such as Wordnet [5], which is an online lexical reference system.

The terms from the vocabulary and their meaning are transparent for the engine which uses only the relations between them. It is the annotator's task to relate the terms, using four different relations: *similar*, *opposite*, *generalization* and *specialization*. Again the annotator, in our case most likely the documentarist, can do that or make use of an existing taxonomy (for example Wordnet uses similar relations).

With this annotation schema Vox Populi is now in the position to generate arguments.

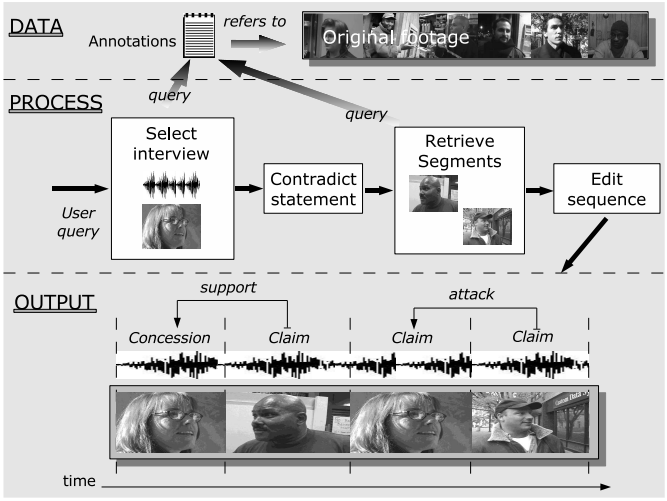

Fig. 2. Example of the Attack strategy referring to the Toulmin structure in Figure 1

5. AUTOMATIC EDITING - GENERATING ARGUMENT LINES

The generation of argumentative progression requires two steps. First, the argument space needs to be created and, second, the line of arguments within this space needs to be generated. The Vox Populi engine first creates a semantic graph that contains the relevant statements as nodes and relations between them as edges, as explained in section 5.1, and then it examines this graph based on rules that inspect the argumentation structures described by the Toulmin model, as explained in section 5.2.

5.1. Creating the argument space

The engine takes each statement encoded as explained in 4.1 and forms related statements in the following way: one at a time the terms in the different parts of the statement are replaced by terms that are related to them. For example, the term `War` is related through *similar* with the term `Military Action`. `War` is also related through *opposite* to the term `Diplomacy`. Based on these structures the engine can generate from a statement, such as "War best solution", the two following statements: "Military Action best solution" and "Diplomacy best solution". This process is repeated for each relation (thus also for *specialization* and *generalization*) and it is not only applied to the initial statement, but also to the statements derived at each step, so that different parts of the statement can be replaced. The rationale behind this process is that the relation between two terms can be used to infer the relation between two statements that contain those terms. In the above-mentioned example, the statement "Military Action best solution" is assumed to support the statement "War best solution", while the statement "Diplomacy best solution" contradicts it. A generated example, using the IWA material, is shown in Table 1.

s	m	p	**operation**
I	not	afraid	
			apply *generalization* on s
People	not	afraid	
			apply *opposite* on m
People		afraid	
			apply *similar* on p
People		threatened	

Table 1. Example of generating three statements from a given one

When all derived statements are generated from the original one, the engine queries the annotation repository to see whether these statements exist. Each hit is linked to the original one, recording also how the statement was derived (in the example in Table 1 the edge linking the two statements is: *generalization* s - *opposite* m - *similar* p). If the statement was derived using no or an even number of *opposite* relations, we assume that it supports the original one. However, the more relations are used to derived a statement, the less we can rely on the above-mentioned conclusion.

5.2. Generating the argument line

In this phase the engine selects the sequence to display to the user. The engine uses one or more strategies to select the content, where each strategy contains a policy to traverse the semantic graph and criteria to select a node or not. The traversal is based on the fact that the edges of the semantic graph are typed. We implemented two strategies, but for reason of space in the following we will only describe the "Support or Attack a Position" strategy. This strategy selects an interview and presents it trying to support or attack the position expressed in it. The latter case can be seen in Figure 1, where the system uses the Toulmin-based description of the selected interview, with the following rule: select video segments expressing supporting statements for the concession and condition, select video segments expressing contradicting statements for the claim, warrant, backing and grounds.

As already mentioned in section 5.1, the edge between two statements determines whether they support or contradict each other: if the edge contains no or an even number of *opposite* relations, then they support each other (e.g. "Bombing not effective" and "War not effective"), while if it does they diverge from each other (e.g. in the example in Table 1). The process is shown in Figure 2.

6. CONCLUSIONS AND RELATED WORK

The described rhetoric processes, which aim to use minimal semantic annotations, demonstrate the feasibility of our approach. The prototype engine is able to generate video sequences with an argumentation structure.

Our approach is similar to [6], where video sequences are annotated with keywords, and the keywords are related to each other. Keyword annotation is less time-consuming but it does not allow to create a semantic graph of video segments linked by typed relations as needed by automatic video generation strategies of the kind introduced in section 5.2. A similar approach of generating a presentation based on traversing a semantic graph is [7], but in that case the semantic graph was given as input to the process, while in our case it is generated from the annotations. Because of this, the documentarist cannot foresee all the possible video sequences generated by the system. The making of this kind of progressing documentaries could change the way documentarists work, especially if they can use tools that facilitate the creation of annotations. This would require a new production environment where the annotations become part of the working process.

Other systems use argumentation relations, or more in general, discourse structure relations in annotations. In [8] the authors exploit the context created by such annotations to retrieve documents in the COLLATE[3] system, a collaborative environment containing historic documents about last century European films.

Our usage of argumentation structure allows the engine to build arguments based on the purpose to support or contradict a particular opinion. In this respect, our work is similar to Terminal Time [9], which creates historical video documentaries that are biased by the audience interacting while the documentary is projected. The major difference between the two approaches is that Terminal Time is content driven, where our approach is structure oriented, which makes it potentially applicable to more domains as discussed in the Future Work section.

7. FUTURE WORK

We believe that our framework can be applied to different documentary types and we are, therefore, investigating how domain dependent our argument space generation in 5.1 is and how to isolate domain dependencies in the system so that the overall architecture can still be applied to different domains. Currently this investigation is carried out within the scope of another video documentary project, Montevideo's Visual Jockey[4].

[3] `http://www.collate.de/`

[4] Available at `http://www.montevideo.nl/en/onderzoek/projecten/vjcultuur.html`, this project describes the work of VJs with respect to other art disciplines and the reciprocal influences between existing visual arts and VJ.

Acknowledgments

This research was funded by the Dutch national ToKeN2000 I2RP and CHIME projects. The authors wish to thank Lloyd Rutledge and Arjen de Vries for useful discussions during the development of this work.

8. REFERENCES

[1] Stephen Toulmin, Richard Rieke, and Allan Janik, *Introduction to Reasoning*, MacMillan Publishing Company, 2 edition, 1984.

[2] Trevor J.M. Bench-Capon, "Specification and Implementation of Toulmin Dialogue Game," in *Proceedings of JURIX 98*, 1998, pp. 5–20.

[3] Marc Davis, *Readings in Human-Computer Interaction: Toward the Year 2000*, chapter Media Streams: An Iconic Visual Language for Video Representation., pp. 854–866, Morgan Kaufmann Publishers, Inc., 1995.

[4] F. Nack, *AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing*, Ph.D. thesis, Lancaster University, 1996.

[5] Sergey Melnik and Stefan Decker, "Wordnet RDF Representation," http://www.semanticweb.org/library/, 2001.

[6] G. Davenport and M. Murtaugh, "ConText: Towards the Evolving Documentary," in *ACM Multimedia '95, Proceedings*, November 1995, pp. 377–378.

[7] Joost Geurts, Stefano Bocconi, Jacco van Ossenbruggen, and Lynda Hardman, "Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations," in *Second International Semantic Web Conference (ISWC2003)*, Sanibel Island, Florida, USA, October 20-23, 2003, pp. 597–612.

[8] Ingo Frommholz, Ulrich Thiel, and Thomas Kamps, "Annotation-based Document Retrieval with Four-Valued Probabilistic Datalog," in *Proceedings of the first SIGIR Workshop on the Integration of Information Retrieval and Databases (WIRD'04)*, July 2004, pp. 31–38.

[9] Michael Mateas, "Generation of Ideologically-Biased Historical Documentaries," in *Proceedings of AAAI 2000*, July 2000, pp. 36–42.