Supporting the Generation of Argument Structure within Video Sequences

Stefano Bocconi, Frank Nack and
Lynda Hardman*
CWI
P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

ABSTRACT

We describe our approach to the automatic generation of argument structures in the domain of video documentaries. Our approach releases control of the final video sequencing from the film maker/annotator to the system and thus allows users to select their own documentaries for viewing. Each video segment is annotated using a formal structure filled in with terms from a thesaurus. The annotations are used for finding and combining video segments into a final presentation. In order to influence the documentaries that can be generated, we introduce three methods for the annotator to evaluate the effectiveness of the annotations and to influence the process of automatic link generation.

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, Argument Structure

1. INTRODUCTION

Vox Populi is a rhetoric engine for automatically generating argumentation video sequences from a semantically annotated media repository. This type of engine supports artistic initiatives, such as Montevideo's VJ project[1] or Interview with America[2] (IWA) where the documentary material is made available to the general public in the form of a documentary space, rather than a single fully-edited linear film.

In authoring a documentary space, the author controls the design of the semantic content through media gathering and media annotation. The latter covers the description of relevant concepts from the domain the documentary deals with, as well as typed relations between the concepts. The author designs the space without having to specify explicitly how and in what order the audience should access the material. This is handled by the rhetoric engine, while the choice of annotations and relations gives the author artistic freedom to provide a particular view or views of the material. The way the material is annotated defines the points of view that the engine can generate.

The challenge for the author is to describe the material in such a way that the rhetoric engine can exploit the potential meaning covered in the semantics. Since the author has no influence on the run-time generation of the video sequences, the expected richness of the generated sequences may not occur, and the documentary system can exhibit poor behavior with respect to the information selected or its presentation. The reasons for this may be that the author made errors when annotating the material, either by introducing inconsistencies of content descriptions or by establishing erroneous relationships between concepts. The author may have forgotten to establish relationships between concepts or simply used a poor set of semantics that does not appropriately express the content of the media items. Whatever the reason, the author can be supported by defining mechanisms that identify the source of the problem and suggest improvement strategies.

In this paper we address this problem, and describe how Vox Populi provides an empirical method that facilitates the author's verification of the documentary space. Specifically, the proposed method provides a means of evaluating the effectiveness of the relations between terms used in the annotations.

We first briefly introduce the Vox Populi presentation engine to establish the framework in which an author designs. We then discuss our approach in the context of related work.

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

HT'05, September 6–9, 2005, Salzburg, Austria.

[1] http://www.montevideo.nl/en/
[2] http://www.interviewwithamerica.com/documentary.html

The sections that follow describe the annotation structures and how the engine operates on them. We then explain our method to support the authoring of a documentary space. We evaluate our approach based on data collected during the design of the "Interview with America" documentary space. The paper concludes with some discussion and future work.

2. THE VOX POPULI PRESENTATION ENGINE

In this section we briefly introduce our rhetoric-driven presentation engine Vox Populi, which is described in [4]. The engine utilizes an audio-visual repository to automatically generate short video sequences that make a point and show argumentation progression. The repository we use, "Interview with America", contains approximately 8 hours of video interviews. United States residents from different socio-economic groups were interviewed on the events happening in the aftermath of the terrorist attack on September 11th, 2001. Issues discussed include the war in Afghanistan, anthrax, media coverage and social integration in multicultural societies.

Vox Populi utilizes two types of annotations: descriptive and rhetorical. The descriptive annotations cover the who, where, when and what in the video. The rhetoric annotations are based on the verbal information contained in the audio channel, identifying the claims the interviewees make and the argumentation structures they use to make those claims. To encode the argumentation structures we use the Toulmin Model [19] which is well established in the literature and commonly adopted. Our approach, however, does not depend on this specific model and could use any argumentation model, providing that it describes the role of the different statements in making a claim.

Vox Populi generates meaningful video sequences by selecting and ordering video segments using rhetoric-based strategies, such as opposition and similarity. Those strategies traverse the graph of typed relations between video segments (the *Semantic Graph*) deriving video sequences from this structure. In Section 3 we introduce other systems that use a semantic graph to generate presentations. For all this systems the content and variety of the presentations they can generate depend on the "richness" of their knowledge base, i.e. the semantic graph. In our case the *Semantic Graph* is the product of the automatic link generation process using the annotation schema. That is why it is important that the annotations are correctly crafted for the automatic linking process to produce a sufficiently rich graph.

In Figure 1 we show an example scenario: the user asks Vox Populi to show interview segments containing contrasting opinions about the war in Afghanistan, with a bias towards people who are against it[3]. The engine first selects an interview which is in favor of the chosen subject (the woman on the top right of Figure 1 saying: "I am not a fan of military actions, but in the current situation I cannot think of a more effective solution"). The rhetoric annotations for this statement decompose it into two parts, the Claim ("I cannot think of a more effective solution") and the Concession ("I am not a fan of military actions").

To contrast her point of view, the engine chooses to support the Concession and contradict the Claim: for the former

[3]A demo of our engine and implementation details can be found at `http://www.cwi.nl/~media/demo/IWA/`

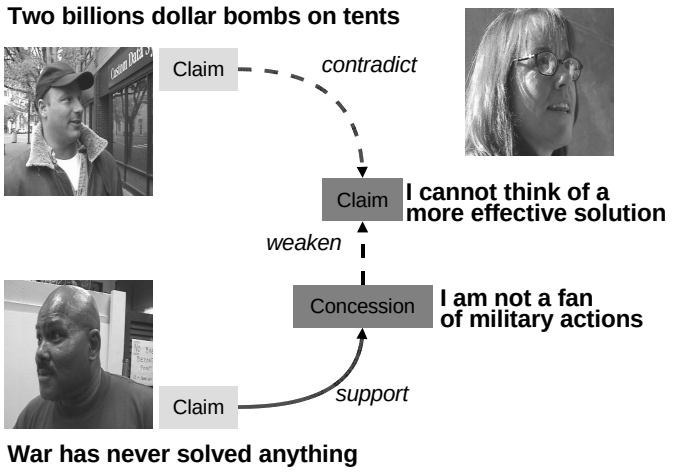

Figure 1: Vox Populi: assembling contrasting points of view about the war in Afghanistan

it selects the man on the lower left saying "War has never solved anything" and for the latter the man on the upper left saying "They are using two billion dollar bombs on ten dollar tents". As explained in [4], the rationale behind this choice is that according to the argumentation model we use, the Concession is a rebuttal of the Claim, i.e. it contradicts the claim even though it is less strong than the claim (according to the claimer). Supporting the Concession makes the rebuttal of the claim stronger. Contradicting the Claim constitutes another form of rebuttal.

The final video sequence is as follows: woman saying "I am not a fan of military actions"; lower man saying "war has never solved anything"; woman saying "in the current situation I cannot think of a more effective solution"; upper man saying "two billion dollar bombs on tents".

As noted above, the capability of Vox Populi to find and select appropriate statements to support or contradict the initial statement is directly related to the quality of the *Semantic Graph*.

3. RELATED WORK

We compare this work with a number of foci of hypertext research, such as semantic-networked discourse, linking and the automatic generation of links, structural analysis and conceptual modeling of semantic navigation structures and conceptual hypermedia.

The notion of a semantic-network based discourse has always been an objective within hypertext research. We have seen developments on modeling argumentation discourse in general [7], sophisticated requirements for scholarly argumentation [20], and on establishing large narratives [21]. This research direction is, despite its linguistic focus, inspirational as it investigates the impact of argumentation on directed graph models, where nodes and links have a semantic structure that can be used to support accessing and visualizing the established or emerging network. Applying access to material based on the connections between concepts, where the connections are grounded in a discourse and/or argumentation ontology, is also the approach in Vox Populi but we go a step further. We do not apply this tech-

nique to present an existing discourse but to generate a new argument flow on-the-fly, depending on the information need of the visitor of the documentary space. In this respect work by Mateas [16] and Davenport and Murtaugh [10] is closer to our own.

The ConText system [10] shares the aim of Vox Populi to allow the author to gather material in an evolving environment. The author, however, is removed from the complex task of explicitly sequencing the material for each viewer. The major difference between the two systems is that in ConText video sequences are annotated with keywords, where the keywords are related to each other. Keyword annotation is less time-consuming but it does not support the creation of a semantic graph of video segments linked by typed relations as needed by automatic video generation strategies of the kind used by Vox Populi [4]. Terminal Time [16] applies rhetoric strategies to generate cinematic experiences for mass audiences. The major differences between it and Vox Populi is that Terminal Time is plan-driven and places media items into the generated rhetoric plan, while Vox Populi is data driven and creates a plan based on the available media items.

Linking plays an important role, a core theme in hypertext literature [14, 18, 23]. Authors can be provided with a collection of techniques for reducing undesired structural complexity and create documents that readers can traverse more easily [3, 5, 13]. Yet the problem of the author in Vox Populi is different since here the author designs the semantic description of the material and leaves the final linking to the system. Nevertheless, the repair mechanisms described in [5] can be part of such automatic linking. The advantages of automatic link generation are discussed in [8, 22]. In the former, Cleary and Bareiss argue about the difficulty of creating a coherent question answering hypermedia system, in terms of time spent by experienced annotators and in errors made when linking parts of textual stories. The authors claim advantages of automatic linking over manual linking, since the former is capable of dealing with changing information. This overcomes one of the limitations of static hypermedia systems (see issue 3 in [12]). Cleary and Bareiss' approach is similar to our own because we both try to capture the aspects of documents which would cause an indexer to create a link between them. Their point linking technique, in particular, is very similar to our annotation structure, although we use exclusively a thesaurus rather than linking rules. Our approach is different because we adopt the point linking ideas but use them for generation as well as for evaluation purposes, as shown later in this paper.

The idea of traversing a semantic graph to generate a presentation of aspects contained in the information space is not new [11, 15]. The difference between Vox Populi and these systems is that in Vox Populi the semantic graph is generated automatically by the engine and not given as input, as in [11]. In [15] the annotations used are neither taxonomic nor structured and the relations built are not argument based.

Finally, conceptual navigation [9] and conceptual hypermedia [6] discuss the use of taxonomies or ontologies to support browsing of the annotated information space. Vox Populi differs from their approach because a thesaurus is used not only to annotate (or classify [6]) information items but also to establish semantic relations between them.

Having introduced the Vox Populi environment and compared it with existing work we will now explain the basic concepts the engine operates on, namely the annotation structures (Section 4) and the generation process (Section 5).

4. THE STRUCTURE OF THE ANNOTATIONS

The quality of the semantic annotations of the media items is vital for the construction of rich argument structures. In this section we explain the basic annotation structures in Vox Populi and describe the process that automatically links different annotations, illustrated by the example in Figure 1. In Vox Populi the statement is the entity that represents the content of a video segment.

4.1 The Statement

In previous work [4] we describe the rhetoric annotations of the verbal information contained in the audio channel of a video interview. These are based on statements, which are intended to capture the semantics of a claim an interviewee makes, for example, "They are using two billion dollar bombs on ten dollar tents".

A statement in Vox Populi is composed of a `subject`, a `modifier` and a `predicate`. The `subject` represents the subject of the statement, the `predicate` qualifies the `subject` and the `modifier` modifies the relation between the `subject` and the `predicate`. The statement "Two billion dollar bombs on tents" in Figure 1, for example, is encoded as `subject:Bombing modifier:not predicate:effective`. Note that this choice of a three-part structure is a trade-off between expressiveness and computational complexity — our findings are not limited to three-part statements.

Each statement corresponds to one or more video segments. Analogously, a video segment can have multiple statements associated with it in case more meanings can be applied to it. A fundamental issue in associating statements with media items is that each video segment must provide enough context to the viewer to assess that the statement applies to what the interviewee has stated. An example of when this does not apply is when interviewees reply to questions with short answers such as "No" or "Yes". Such video segments do not offer enough context to convey what the intention of the interviewee is. Their meaning depends entirely on what is shown before or after them. Although potentially interesting, our approach does not (yet) make use of such video segments.

To give an idea of the order of magnitude we are dealing with, one hour of video from the IWA material has been annotated resulting in 118 encoded statements using 155 terms for the three parts of a statement. The terms are contained in a thesaurus, as explained below.

4.2 The thesaurus

Each term used as a value in one of the three parts of a statement belongs to a thesaurus, which is also composed of three parts, one for each of the three parts of a statement. A `subject` can thus not have the same value as or be related to a `modifier` or `predicate`. 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, as explained in Section 5.

In Vox Populi, a documentarist (turned annotator) can build a vocabulary in parallel to annotating media items.

	Bombing	War	Peace	Diplomacy	Military Actions	Economic Aid
Bombing	*Id*	*Gen*	— —	— —	— —	*Opp*
War	*Spec*	*Id*	— —	*Opp*	*Sim*	*Opp*
Peace	— —	*Opp*	*Id*	— —	— —	— —
Diplomacy	— —	*Opp*	— —	*Id*	*Opp*	— —
Military Actions	— —	*Sim*	— —	*Opp*	*Id*	— —
Economic Aid	*Opp*	*Opp*	— —	— —	— —	*Id*

Table 1: Example of relations between terms for the `subject` thesaurus

	no modifier	not	never
no modifier	*Id*	*Opp*	*Opp*
not	*Opp*	*Id*	*Sim*
never	*Opp*	*Sim*	*Id*

Table 2: Example of relations between terms for the `modifier` thesaurus

	effective	waste	useless
effective	*Id*	*Opp*	— —
waste	*Opp*	*Id*	*Sim*
useless	— —	*Sim*	*Id*

Table 3: Example of relations between terms for the `predicate` thesaurus

She can instantiate statements based on her own vocabulary or can make use of an existing thesaurus, such as Wordnet [17]. In either case, a vocabulary for the particular documentary space is created. For example, to create the statement in Section 4.1 `Bombing not effective` each term must be present in the thesaurus and related to other terms in it.

Vox Populi requires that the terms used in any of the three parts are related using four different relations: similar (hereafter *Sim*), opposite (hereafter *Opp*), generalization (hereafter *Gen*) and specialization (hereafter *Spec*). These relations correspond to the canonical relations in a thesaurus: *synonym*, *antonym*, *hypernym* and *hyponym*, respectively. The annotator, in our case most likely the documentarist, can create these relationships explicitly among the terms in each of the three sub-thesauri, or can make use of an existing thesaurus (for example, Wordnet also uses *synonym*, *antonym*, *hypernym* and *hyponym*).

In Tables 1, 2 and 3 we show example terms and the relations between them (as <*row*><*relation*><*column*>) from the `subject`, `modifier` and `predicate` thesauri. In our case the relations *Opp* and *Sim* are symmetrical, while *Gen* is the inverse of *Spec*; nevertheless, our approach neither depends on this nor it requires it.

In defining the annotations and the thesaurus, the author is free, using the annotations, to establish the particular semantics of the documentary space. One of the most important issues for the author is, therefore, to establish whether the domain represented by the collection of media items is covered by the vocabulary developed, i.e. whether the specified terms and relations in the thesaurus describe the media items sufficiently to make the content available to the end-user.

For example, in Table 1 `Diplomacy` is *Opp* to `War`, but not to `Bombing`. The decision not to relate two terms has consequences for the linking process, as explained in Section 5, but the annotator at authoring time does not have any insight on how her decisions are influencing that process.

To give an idea of the order of magnitude we are dealing with, the total number of relations defined in the thesaurus is 199.

Before we describe the mechanisms to support the author in these aspects of documentary space design, we first briefly outline in Section 5 the process of manipulating statements to automatically relate media items. This mechanism is used by the generation engine to establish the relevant story space in the form of a *Semantic Graph*, which can then be traversed with the appropriate rhetorical strategies according to the information need of the audience. As shown later, this step is relevant for supporting the author as it can be utilized to establish the quality of the thesaurus.

5. AUTOMATIC LINK GENERATION

The aim of the Vox Populi engine is to automatically assemble short meaningful video sequences with an argumentation progression that represents a particular information request by a user. The engine starts with an opinion (expressed by one or more statements, as shown in Section 2) and, by manipulating it, locates those statements in the repository that fulfill the argumentation progression requested by the visitor. Juxtaposing the statements, through the corresponding video sequences, gives the requested output.

The crucial part of this process is the generation of a *Semantic Graph* structure based on the annotations described in Section 4.1. The graph represents the relationships between statements and thus spans the potential conceptual space for the documentary sequence to be produced. The process consists of two sub-processes: deriving new statements from existing ones and verifying whether these new statements are present in the annotations in the repository.

5.1 Deriving New Statements

In the first phase of the process the structure of the statement, as explained in 4.1, is taken as the basis for the graph generation. One at a time the three parts of the statement are replaced by terms that are related to them. For example, the subject `Bombing` might be associated through the relation *Opp* with the subject `Economic Aid`. `Bombing` is also related through *Gen* to the subject `War`. The engine can thus generate, from the statement `Bombing not effective`, the two following statements: `Economic Aid not effective` and `War not effective`. This process is repeated for every part of the original statement using the relations in the thesaurus.

Statement			Steps	Video Present
subject	**modifier**	**predicate**		
Bombing	*no modifier*	effective	Modifier *Opp*	Yes
War	not	effective	Subject *Gen*	Yes
Economic Aid	not	effective	Subject *Opp*	**No**
War	*no modifier*	effective	Subject *Gen*, Modifier *Opp*	Yes
Bombing	*no modifier*	waste	Predicate *Opp*, Modifier *Opp*	Yes
Peace	not	effective	Subject *Gen*, Subject *Opp*	**No**
Diplomacy	not	effective	Subject *Gen*, Subject *Opp*	Yes
Military Actions	*no modifier*	effective	Subject *Gen*, Modifier *Opp*, Subject *Sim*	Yes
Bombing	*no modifier*	useless	Predicate *Opp*, Predicate *Sim*, Modifier *Opp*	Yes
Bombing	possibly	effective	Modifier *Sim*, Modifier *Opp*, Modifier *Sim*	**No**
War	not	useless	Subject *Gen*, Predicate *Opp*, Predicate *Sim*	**No**
Economic Aid	*no modifier*	effective	Subject *Gen*, Modifier *Opp*, Subject *Opp*	Yes
Military Actions	never	effective	Modifier *Sim*, Subject *Gen*, Subject *Sim*	Yes
Peace	once	effective	Modifier *Sim*, Subject *Gen*, Modifier *Opp*, Subject *Opp*	**No**
Hate	never	effective	Modifier *Sim*, Subject *Gen*, Subject *Opp*, Subject *Opp*	**No**

Table 4: Example of statements generated from `Bombing not effective` (column 1), the steps in the process to generate them (column 2) and whether they are present in the repository (column 3)

In addition, each derived statement is also transformed, so that the parts of the original statement are replaced multiple times.

An example of the results of this phase is shown in Table 4. Here, for example, the sixth statement, `Peace not effective`, has been generated from the statement `Bombing not effective` in two steps. First by using the relation *Gen* between `Bombing` and `War` giving `War not effective` (second row in Table 4), and then by using the relation *Opp* between `War` and `Peace` (see also Table 1 for the relations between terms).

At the present stage we are generating new statements applying the relations in the following order: *Sim*, *Gen*, *Spec* and *Opp* and we iterate this step twice. We will discuss this choice in Section 8.

While all the statements generated are well-formed, not all of them exist as annotations in the repository, because some of them have no corresponding media item. These are identified as the 'No' items in the last column of Table 4. Note, the proportion between present and absent statements in the table does not reflect the usual proportion found. The majority of generated statements is normally not present, as explained in Section 7.1.

Referring to the missing relation between `Bombing` and `Diplomacy` discussed in Section 4.2, the process is able to relate statements containing those terms via the chain of substitutions `Bombing` ⇒ `War` ⇒ `Diplomacy`, as can be seen in Table 4 (a statement containing `Diplomacy` is generated in two steps, seventh statement). If the thesaurus did not contain a relation between `War` and either `Bombing` or `Diplomacy`, the statements `Diplomacy not effective` and `Bombing not effective` would not be related.

The result of this first phase is a graph of statements (the nodes) connected by typed relations (the edges).

5.2 Querying for New Statements

The goal of the second phase of the process is to select only the statements (nodes in the *Semantic Graph*) that correspond to media items in the repository. In others words, this phase transforms a *Semantic Graph* of well-formed annotations into a *Semantic Graph* of media items contained in the repository. This is done by querying the annotations repository and simply eliminating the nodes from the graph that have an empty result set. Note that a result set can contain multiple hits for the same statement because several media items can be annotated with the same statement. The 'Yes' terms in the last column of Table 4 indicate that the generated statement is present in the repository and corresponds to at least one video segment.

0	**1-4**	**5-8**	**9-12**	**13**
54	47	4	4	9

Table 5: Number of statements having number of links in x-y range (13 being the max number of links)

The end result of the linking process is a *Semantic Graph* of related statements where each is guaranteed to have a corresponding video segment in the repository. The relations are exploited by the rhetoric engine to establish the final argumentation structure for the sequence to be presented.

6. QUALITY IS QUANTITY

As shown, the author can concentrate on the task of designing the semantics of a documentary space through media gathering and media annotation, leaving to the system the time-consuming task of specifying how and in what order the audience should access the material. This freedom is not unproblematic, however. The problem is that the quality of the provided annotations and related thesaurus is directly related to the quality of the *Semantic Graph*. From an authoring point of view it is crucial that the annotations and the thesaurus are correctly crafted for the automatic linking process to produce a sufficiently rich graph, i.e. a graph whose exploitation provides access to sufficient media items.

In the case of manual linking, the author/annotator can have an idea about how connected the documents in the repository are—this is not true for automatic linking. The questions an author in a Vox Populi environment is con-

Min	Max	Average
1	1610	203,1

Table 6: Result for the first index defined in Section 7.1

fronted with are, for example: "Can I generate a presentation about every topic contained in the repository?", "Is every video segment contained in at least one presentation?" and "Do I have to worry about the fact that presentations are too long or too short?". These questions are especially relevant considering that our approach is potentially able to make all the material contained in the repository available to the viewer, avoiding the possible information loss caused by a final version. This requires that the automatic linking process actually links all the video segments that a human author would consider to be related.

An author can benefit from help with verifying whether the created semantic space is efficiently descriptive and connected, as mentioned above. While working on the IWA environment, for example, it turned out that the rhetoric engine was producing insufficient arguments for the number of statements in the annotation repository. An analysis of the linking process showed that 54 statements out of 118 were not linked, and were effectively lost for the aim of the IWA space, because video segments are selected to form a presentation following the links in the *Semantic Graph*. Table 5 gives an overall view on the links in the repository.

When observing the statements that were not linked, we found that some of them were correctly not linked because their semantics was not contained elsewhere in the repository (e.g. the semantic content "Israel is a secure country" was only present in one statement). On the other hand, statements such as "Peace is possible" or "War exceptional" should have been linked, since other similar statements were present in the repository.

Based on this observation, we decided to provide an empirical method to analyze the complete *Semantic Graph*, in order to facilitate an author in evaluating the effectiveness of the semantics defined by the annotations. Such an analysis can point out where and why the automatic linking process is not performing as expected, as explained in the following sections.

7. RE-ENGINEERING THE PROCESS

The measurements described in this section facilitate the evaluation of the performance of the automatic linking process. Their interpretation allows the author to see how effective her annotation structure is in creating a "powerful" *Semantic Graph*, in the sense specified in Section 6.

When the automatic linking process is not capable of connecting sufficient statements in the repository, two solutions (not mutually exclusive) are possible:

1. Modify the annotations of the video segments
2. Modify the relations in the thesaurus

The first solution involves re-annotating the video segments whose statement is not connected to (or insufficiently connected to) other statements in the *Semantic Graph* (as pointed out in Table 5). This approach can solve local problems, but it has two disadvantages. First, it may not be feasible to change potentially many hours of annotating work.

Min (%)	Max (%)	Average (%)	Best Percentage
0 (0%)	13 (3,9%)	2,3 (1,1%)	8,5%

Table 7: Results for the second index defined in Section 7.1, as value and as percentage of the generated statements

Second, the annotator has no guidance on how to change the annotations and might increase the problem by introducing terms with even less connecting potential, resulting in fewer statements being linked.

More promising is to perform changes on the way the terms are related in the thesaurus, as that can have an impact on the performance of the linking process. Our qualitative measurement mechanisms, based on the IWA repository and thesaurus, focus on supporting this type of modification. We show how the supporting measurements are applied to the linking process (Section 7.1) and to the relations in the thesaurus (Section 7.2).

7.1 Measuring Performance of the Statement Generation

In this section we introduce two performance indexes related to the generation process described in Section 5, namely:

1. The number of statements generated from a single statement (process described in Section 5.1).
2. The percentage of generated statements present in the repository with respect to the generated ones (process described in Section 5.2).

These indexes are calculated for each statement. For the sake of clarity, if we apply the indexes to the example in Table 4, assuming that the table presents all statements that can be generated from `Bombing not effective`, the index for the first measurement would result in 15 (the number of statements), while the second index would result in the value $9/15 = 60\%$ (9 being the number of statements with "Yes" in the last column).

The reason to introduce the first index is that the lower the number of generated statements, the less likely it is to find some of them in the repository. The index is thus related to the probability of a statement to be linked to others in the repository.

Considering that the generation of new statements depends on the relations defined in the thesaurus (as explained in Section 5.1), the first index can give an idea about how well the terms in the thesaurus are related to each others. A low value of the first index informs the author that terms in the statement are not well connected in the thesaurus. This can be intentional but could also indicate that the author might have overseen potential relations for the terms. The index gives thus a suggestion to review those terms in the thesaurus.

We (in the role of the author) applied this index to the IWA repository and the results are shown in Table 6 (please note that the tables show summaries, the engine produces detailed information per statement). The most important data in the table is the minimum number of statements generated (column 3), which is 1. This is very low and indicates that there is at least one statement containing terms that are insufficiently related in the thesaurus. The statement with

index = 1 is the following: `Daily Life partially changed`. This is actually one of the statements with zero links in Table 5. When we verified the terms in the thesaurus, we found that we had left out possible relations, since in the repository statements were present that contained for example the subject `Daily Life`, namely `Daily Life changed` and `Daily Life normal`. It can be argued that a small repository such as IWA (118 statements) will necessarily have semantically isolated statements, but as the above case showed, the index served as an indicator for a problem, which the author can ignore or consider. In the above case we considered the warning and added the overlooked relations.

Generating sufficient statements from the original claim increases the probability of constructing an argument, but more importantly, as shown earlier, is to generate those that can be found in the repository. For an author it is important to know how well a statement functions in this respect. This is the reason for the second measurement index we introduced.

In Table 7 we describe some general results based on the second index applied to the IWA repository. We discuss here only the minimum and maximum values, while we will be able to examine average and best percentage (reported here for completeness) when we have data about other repositories.

The results for the minimum value, namely 0, could be expected as the value of 54 statements in Table 5 already showed that these are semantically isolated items. Yet, beyond this information of "unlinkedness", a low value can also indicate that despite a good performance on 'generating' (see first index), the automatic linking performs poor on 'retrieving'. The process described in Section 5.1 might be generating "nonsense", i.e. statements that from the point of view of repository semantics do not make sense and are not present. For example, the statement `People threatened` generates 1342 statements, of which 4 are present in the repository (0,3%). The high number of generated statements (the first measurement index) indicates that the terms are well connected in the thesaurus, but these relations are not able to produce sufficient statements with material contained in the repository. In Section 7.2 we will return to this point again and show how the particular relations that cause this type of bad performance can be pointed out. At this point of our discussion we see that the second measurement index can suggest to the author that there is a gap between the semantics of the thesaurus and the annotations in the repository.

Again, it is the author who reflects and reacts to the suggestions made by the system. For example, the author might decide to consider the low values as suggestions to add annotations to the repository, because there is potential for linking them (i.e. there are many generated statements but only few retrieved).

Note that a low value on the second index is not in itself such a problem. This could be due to the fact that many statements are generated and the repository contains only a few (see for example Table 6, 1610 statements are generated while the repository contains only 118). Nevertheless, it can indicate a problem when 45,8% of the statements (54 out of 118) have no links, as shown in Table 5.

The maximum value in Table 7 gives an indication about what a high value for the second index can be: less than 5% when retrieving the most statements (or 8,5% when considering the best ratio between retrieved statements and generated statements). This provides the author with an upper limit above which she will probably not be able to improve the value of the second index.

We showed so far mechanisms that indicate which statements do not contribute to generating a rich *Semantic Graph*. In the next section we show how to find out relations in the thesaurus that are not capturing the semantics of the repository.

7.2 Measuring Performance of the Thesaurus

In this section we introduce a measurement to identify those relations that cause bad performance of the automatic linking process (described in Section 5). In order to do so, we keep track of the relations used to generate each statement. For example, the sixth statement in Table 4, `Peace not effective`, has been generated using two relations: relation *Gen* between `Bombing` and `War`, and then relation *Opp* between `War` and `Peace`.

If a generated statement is found in the repository, the relations used to generate it caused a "hit", while otherwise the relations caused a "miss". The hit and miss scores form the basis of this performance index.

In the above example (`Peace not effective`), both relations get 1 point on the miss score, since a video segment corresponding to the statement is not present in the repository (as indicated in Table 4).

Calculating these scores for the relations for the IWA thesaurus, we found that out of 199 relations, 101 had a hit score of zero, i.e. they were never able to generate a hit. The process of automatic linking would generate the same links even though these relations would not be present in the thesaurus. The result is a clear sign that these relations are not used to generate the *Semantic Graph*. For the author this means she should consider eliminating these relations (at least as long as the content of the repository does not change) or modifying them.

Table 8 shows the relations with the highest miss score. Among them, the relations with a zero hit score are the ones that can easily be deleted as they only consume computer resources. For the others, the author should consider whether they describe a few but valuable semantic options, which are simply kept, or if they are misconstructed and should be modified.

For example, the first and the fifth row in Table 8 seems to indicate that the author was interested in making the semantic distinction between `Normal People` and `Rich People`, both *Spec* from `People`. This distinction, though, has generated a large number of miss scores (more than 3000) in comparison to 8 hits. This semantics apparently is weakly represented in the repository as far as the annotations are concerned. Recalling that the statement `People threatened` in Table 7 generates many statements but retrieves few (as discussed in Section 7.1), the author can now see the relations causing the low number of hits and thus address the problem.

On the other hand, Table 9 shows the best 10 relations according to the ratio $\frac{hit}{miss}$. These relations are capturing the semantics of the repository and should not be changed, unless the author makes a conscious decision to change the semantics of the repository. For completeness we also report in Table 10 the best 10 relations according to the hit score.

The tools used in our prototype to provide support for the

Term 1	Relation	Term 2	Hit Score	Miss Score
People	*Spec*	Normal People	8	-3802
I	*Gen*	People	10	-3084
no modifier	*Sim*	best	3	-2292
not	*Sim*	never	8	-2253
People	*Spec*	Rich People	0	-2110
fearful	*Sim*	attentive	8	-2049
no modifier	*Sim*	can	0	-1925
Americans	*Gen*	People	5	-1837
no modifier	*Sim*	always	11	-1755
War	*Gen*	Violence	1	-1460

Table 8: Worst 10 relations based on "miss" score

Term 1	Relation	Term 2	Hit/Miss
Economic Aid	*Opp*	Bombing	16%
waste	*Opp*	effective	14%
Diplomacy	*Opp*	War	11,9%
only	*Opp*	not only	16,6%
Economic Aid	*Opp*	War	10,8%
not only	*Opp*	only	15.8%
not best	*Opp*	*no modifier*	10,53%
Ground Forces	*Opp*	Bombing	9,9%
only	*Sim*	not only	8,9%
not only	*Sim*	only	8,3%

Table 9: Best 10 relations based on $\frac{hit}{miss}$ ratio

author during the design of the semantic space were used by the IWA group while developing the IWA environment. So far the empirical data provided by the system was used to improve the thesaurus design as well as the annotation space.

8. DISCUSSION AND CONCLUSION

In this paper we described particular authoring support provided by Vox Populi, a rhetoric engine for automatically generating argumentation video sequences from a semantically annotated media repository. We described the role of the author in this type of system, namely to design the semantics of a documentary space through media gathering and media annotation without having to specify explicitly how and in what order the audience should access the material. We also showed that the design of the annotations as well as the related conceptual space in the form of a thesaurus are challenging. The biggest problem for the author is to distinguish the effectiveness of the created semantic space. We developed and implemented three indexes that facilitate an author in testing the effectiveness of the established semantic space in generating the story space. The mechanisms mainly cover the detection of effective and non-effective relations in the thesaurus with respect to linking the material in the repository. The developed mechanisms were used by the designers of the IWA space.

The described support mechanisms demonstrate the feasibility of our approach. We adopted these mechanisms to provide various sets of empirical data which have been and still are used to improve the IWA thesaurus as well as the annotations of video segments in the IWA repository. The current approach needs, however, further fine tuning.

As we see it, the research presented in this paper presents a natural evolution in two directions, for both of which we have already taken the first steps:

- Automatic link suggestion
- Re-engineering the automatic linking process (not only the thesaurus)

In the first case the engine is able to identify relations in the thesaurus that are not effective, and can also suggest relations to add between particular terms. A way to achieve this is to start with a fully connected thesaurus (thus every term connected to every other term) and measure the index described in Section 7.2, then suggest to retain only the relations that score best or above a certain threshold.

The second research direction involves measuring how effective each iteration of the process described in Section 5.1 is in generating a hit in the repository. We already have some data that relate the number of hits to the number of transformations used to generate the statement causing an hit. We observed that after a certain number of transformations there are no more hits. This seems to indicate that generated statements can become semantically too far from the original content of the repository and the data could help define a limit for the number of iterations to use.

This last issue is particularly important for the computational complexity of our approach, especially as the semantic space to be designed is large. Our approach is tested so far on a medium sized repository. As already said in Section 5.1, we use two iterations and the generation of a *Semantic Graph* takes 90 seconds (including generating the performance indexes). Using 3 iterations the process takes 331 seconds, while with 4 it takes 768. The number of itera-

Term 1	Relation	Term 2	Hit Score	Miss Score
War	*Spec*	Bombing	45	-970
Bombing	*Gen*	War	39	-746
no modifier	*Opp*	not	32	-876
not	*Opp*	*no modifier*	30	-684
Military Actions	*Sim*	War	22	-363
waste	*Opp*	effective	21	-150
War	*Sim*	Military Actions	21	-527
effective	*Opp*	waste	21	-1561
not best	*Opp*	*no modifier*	12	-114
Bombing	*Opp*	Ground Forces	1	-1460

Table 10: Best 10 relations based on "hit" score

tions is thus crucial for the performance of our approach. On the other hand, we believe that performance depends less on the number of statements and our approach should be scalable to environments with a large repository. Performance is more likely to be influenced by the number of terms and relations in the thesaurus. In considering computational complexity it is important to notice that the support we provide is at authoring time and we are not constrained by run-time requirements.

In Section 7.1 we pointed out that the retrieved quality measurements are subject to interpretation and we provided hints on how the data could be automatically interpreted. To date, we can only use the observations gathered during the work with the IWA project. We are also collaborating with another team of a documentary project, namely the VideoJockey (VJ) project from Montevideo[4]. In this context we try to learn more about the flexibility of the rhetoric approach of Vox Populi and simultaneously investigate the ways in which authors establish their semantic space. Our assumption is that applying our approach will allow us to heuristically define threshold values for our performance indexes. We believe that our evaluation mechanism, although developed in the scope of the Vox Populi engine, can be of interest to other approaches where a thesaurus and structured relations are used to infer relations between annotated items.

Another interesting research direction is to apply our approach to collaborative annotation efforts. In this case, where the role of the author is shared among a number of people, inconsistent annotations are more likely, and the use of author support tools such as those presented in this paper could be valuable.

In this paper we discussed issues related to the generation of the *Semantic Graph*, i.e. the story space. This step forms the basis for our approach to automatically generating video documentaries. We are currently researching how to exploit this space with strategies particularly geared to presenting video, looking at disciplines such as film theory and narrative.

Acknowledgments

This research was funded by the Dutch national ToKeN2000 I^{2}RP and CHIME projects. The authors wish to thank in particular their colleagues Jacco van Ossenbruggen and Lloyd Rutledge for useful discussions during the development of this work. The authors would also like to thank the IWA team as well as Annet Dekker from Montevideo in Amsterdam for the insightful collaboration on their documentary work.

[4] http://www.montevideo.nl/en/

9. REFERENCES

[1] ACM. *The Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, Pittsburgh, PA, June 20-24, 1998. ACM Press. Edited by Kaj Grønbæck, Elli Mylonas and Frank M. Shipman III.

[2] ACM. *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*, San Antonio, Texas, USA, May 30 – June 3, 2000.

[3] M. Bernstein. Patterns of hypertext. In *The Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia* [1], pages 21–29. Edited by Kaj Grønbæck, Elli Mylonas and Frank M. Shipman III.

[4] S. Bocconi, F. Nack, and L. Hardman. Using Rhetorical Annotations for Generating Video Documentaries. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) 2005*, July 2005.

[5] R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*, 10(2):142–180, April 1992.

[6] L. Carr, S. Bechhofer, C. Goble, and W. Hall. Conceptual Linking: Ontology-based Open Hypermedia. In *The Tenth International World Wide Web Conference*, pages 334–342, Hong Kong, May 1-5, 2001. IW3C2, ACM Press.

[7] L. M. Carter. Arguments in Hypertext: A Rhetorical Approach. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia* [2], pages 87–91.

[8] C. Cleary and R. Bareiss. Practical methods for automatically generating typed links. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 31–41, Bethesda, Maryland, United States, March 1996.

[9] M. Crampes and S. Ranwez. Ontology-supported and ontology-driven conceptual navigation on the World Wide Web. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia* [2], pages 191–199.

[10] G. Davenport and M. Murtaugh. ConText: Towards the Evolving Documentary. In *ACM Multimedia '95, Proceedings*, pages 377–378, November 1995.

[11] J. Geurts, S. Bocconi, J. van Ossenbruggen, and L. Hardman. Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations. In *Second International Semantic Web Conference (ISWC2003)*, pages 597–612, Sanibel Island, Florida, USA, October 20-23, 2003.

[12] F. Halasz. Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems. *Communications of the ACM*, 31(7):836–852, July 1988.

[13] G. P. Landow. *The rhetoric of hypermedia: some rules for authors*, pages 81–103. MIT Press, Cambridge, MA, USA, 1991.

[14] G. Liestöl. Aesthetic and rhetorical aspects of linking video in hypermedia. In *Proceedings of the ACM European Conference on Hypermedia Technology (ECHT'94)*, pages 217–223, September 18–23, 1994, Edinburgh, 1994. ACM, ACM Press.

[15] S. Little, J. Geurts, and J. Hunter. Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing. In *6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 158–189. Springer, September 2002.

[16] M. Mateas. Generation of Ideologically-Biased Historical Documentaries. In *Proceedings of AAAI 2000*, pages 36–42, July 2000.

[17] S. Melnik and S. Decker. Wordnet RDF Representation. http://www.semanticweb.org/library/, 2001.

[18] S. P. Tosca. A Pragmatics of Links. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia* [2], pages 77–84.

[19] S. Toulmin, R. Rieke, and A. Janik. *Introduction to Reasoning*. MacMillan Publishing Company, 2 edition, 1984.

[20] V. Uren, S. B. Shum, G. Li, J. Domingue, and E. Motta. Scholarly Publishing and Argument in Hyperspace. In *The Twelfth International World Wide Web Conference*, pages 244–250, Budapest, Hungary, May 20-24, 2003. IW3C2, ACM Press.

[21] J. Walker. Piecing together and tearing apart: finding the story in afternoon. In *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*, pages 111–117, Darmstadt, Germany, February 21-25, 1999. ACM. Edited by Klaus Tochterman, Jorg Westbomke, Uffe K. Will and John J. Leggett.

[22] R. Wilkinson and A. F. Smeaton. Automatic link generation. *ACM Computing Surveys*, 31(4):27, 1999.

[23] P. Zellweger, B.-W. Chang, and J. D. Mackinlay. Fluid Links for Informed and Incremental Link Transitions. In *The Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia* [1], pages 50–57. Edited by Kaj Grønbæck, Elli Mylonas and Frank M. Shipman III.