Creating a semantic-based discourse model for hypermedia presentations: (un)discovered problems

Kateryna Falkovych
Centrum voor Wiskunde en Informatica
P.O.Box 94079
NL-1090 GB Amsterdam, the Netherlands
+ 31 20 592 4216
Kateryna.Falkovych@cwi.nl

Stefano Bocconi
Centrum voor Wiskunde en Informatica
P.O.Box 94079
NL-1090 GB Amsterdam, the Netherlands
+ 31 20 592 4202
Stefano.Bocconi@cwi.nl

ABSTRACT
We address the problem of building a flexible discourse model that enables creation of relatively complex discourse structures by using a semantic framework and semantic descriptions of media items. With such model we intend to support different approaches to hypermedia presentation authoring. We give a short overview of existing hypermedia presentation generation systems and discuss what type of authoring they support. We present our motivation for working towards a discourse model that improves on existing approaches. We describe in more detail our experiences with various strategies for building discourse structures for DISC and Samp*L*e systems. Based on these experiences we come up with a set of requirements for a common discourse model.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous; H.5.1, H.5.4 [**Information Interfaces and Presentation:**]: Hypertext/Hypermedia - architectures; Multimedia Information Systems - Hypertext navigation; I.7.2 [**Document and Text Processing:**]: Document Preparation - Hypertext/hypermedia, Markup languages, Multi/mixed media, standards

General Terms
Discourse, Model, Structure, Domain

Keywords
Discourse modeling, Discourse structures, Semantics, OWL, RDF, Hypermedia presentation

1. INTRODUCTION
One of the problems addressed by hypertext research is how to build meaningful navigation structures based on a particular type of discourse relationships between elements. There are various approaches that model discourse in general [3], sophisticated requirements for scholarly argumentation [10], or establishing large narratives [11]. Moreover, research is performed on conceptual hypermedia [2] that discusses the use of taxonomies or ontologies to support browsing of the annotated information space. Though all these approaches model discourse on quite detailed levels they do not allow the automatic generation of discourse structures based on explicit rules.

The semantic web stimulated the appearance of hypermedia presentation generation systems that model discourse using semantic web technologies. These systems use resources from the web or (shared) repositories as the material for presentations. During the last three years of research at CWI we have developed SemInf [7], Topia [9], DISC [5] and Samp*L*e [4] systems. Research very close to our own includes the Artequakt project [6]. One of the main challenges for these systems has been to provide a mechanism, appropriate schemes and annotations to generate discourse structures for presentations "on the fly". This mechanism should be general enough to be applicable across a variety of media types, providing different content, and a variety of genres. At the same time, the mechanism should be specific enough to support creation of coherent discourse structures. We refer to this mechanism as a discourse model.

We define a *discourse model* as a model that has knowledge about ways of composing various genres, enables building various discourse structures for the same genre with the same or different main character (topic) and allows population of a created discourse structure with media items from a repository.

A *genre* is a distinctive type or category of literary composition, such as tragedy, comedy, novel or story [1]. The literary composition determines the specific features of each category, such as a flow of discourse.

A *discourse structure* identifies the flow of a discourse by specifying what concepts a discourse should talk about and in what order these concepts should be discussed. Additionally, it contains some properties of a document structure such as the division of the discourse into sections and subsections.

When we talk about *order*, we assume a meaningful order that arranges a set of items into a sequence. This order is based on some type of semantic relations between items.

We distinguish a discourse model from a *generation process*, which is used to process knowledge present in a discourse structure to

build a final presentation.

A discourse model is assumed to be declarative, therefore all knowledge it contains can be explicitly represented, reused and exchanged between various applications. Knowledge present in a generation process is procedural and thus might be difficult to update (e.g. due to a chosen type of data structures), to reuse and to exchange (knowledge/rules are hard-coded inside a particular language (e.g. Prolog) and cannot be easily extracted). Therefore, it is expedient to provide a larger part of system heuristics in the declarative part.

This paper aims at describing approaches taken by existing systems for hypermedia presentation generation with an emphasis on the problem of building various discourse structures. In Section 2, along with identifying similarities and differences between approaches, we identify gaps and challenges that are still present or have not yet been addressed. We discuss our experiences with building discourse structures with various approaches for different systems. Based on this discussion we present in Section 3 our motivation to work towards a common discourse model. We also identify a set of requirements that this model has to address. We conclude in Section 4 with our aims for the workshop.

2. HYPERMEDIA PRESENTATION GENERATION ON THE SEMANTIC WEB

In this section we describe a number of hypermedia authoring systems that use semantic web technology as means for producing a hypermedia presentation. For each system we describe the mechanism for creating discourse structures. Existing hypermedia presentation generation systems take different positions along the trade-off scale between the amount of human effort involved and the richness of discourse structures produced. This scale represents in a way a path of the development of semantic web technologies. At the beginning of this path, systems were using RDF as means for describing their resources and processing rules and could create a simple non-narrative hypermedia presentations with a minimum of human effort (Topia [9]). At the next stage systems began to use domain schemes (ontologies) and human-authored discourse templates operating on the domain concepts to drive the development of the story in the presentation (Artequakt [6], DISC [5], Samp*L*e [4]). Our view is that for further improvements of hypermedia authoring the amount of development effort can be reduced by enriching systems with additional heuristics about genre and discourse structure composition principles. These explicitly defined heuristics will enable interoperability on the process level to accompany still developing interoperability on the data level (shared annotated repositories). Our aim is to accomplish the maximum pay-off from the amount of human effort involved while keeping the quality of presentation on the level achieved by manually authored discourse.

Topia [9] aims at creating a hypermedia presentation from a set of media items retrieved as a result of a query. Assuming that a retrieval mechanism returns or is able to filter relevant results, the system tries to present the retrieval results hierarchically in a meaningful order. Topia is operating on an RDF multimedia repository of the Rijksmuseum collection [8]. The repository contains annotated media resources. Retrieval results of a query are clustered using the concept lattice technique. Clusters get created out of media items that have the same value of a certain property in RDF. For example, media items that have Rembrandt as creator could form one cluster. For the final presentation the biggest clusters get selected and ordered according to the decreasing size of the clusters. Alternatively there is a possibility for a user to identify his preferences in particular topics. Then clusters corresponding to those topics will be presented before the others. As the result, a user gets a presentation where media items are arranged based on simple structures. The sequence of the clusters forms hierarchical relationships between them. Media items appear inside each cluster presented with the cluster name. Hierarchy, sequence and ordering are typical structures used in documents and thus are familiar to users.

Summary: The approach is domain-independent to the extent that semantic relations are used in a "numeric" way. Thus, there is no dependency between semantics of the found clusters and their arrangement in the final structure. Accordingly in Topia we talk about a document structure with simple ordering mechanisms that guides presentation construction rather than a discourse structure. Ordering occurs on the level of domain concepts and it does not take their semantics into account.

SemInf [5] system uses Open Archive Initiative [7] repository of media items annotated with Dublin Core (DC) concepts to provide an answer to a user query in the form of a short multimedia presentation. The semantic inference engine consists of the set of rules that infer higher-level semantic relationships between media items from DC annotations. For example, if a text fragment has A. Lincoln as its dc:creator and an image has A. Lincoln as its dc:subject then it is possible to infer that this is an image of the creator of the text fragment. Other examples of inferred relations include precedes, follows, describes. These higher-level semantic relations allow grouping of semantically related media items. The final ordering of (groups of) media items into a presentation is achieved by exploiting chronological relationships where possible. To represent the resulting structure visually the higher-level relations are mapped into spatio-temporal relations of the presentation.

Summary: The set of rules SemInf provides is domain independent since Dublin Core includes a very simple schema applicable across domains. There is no mechanism defined that guides the creation of a discourse structure. Ordering occurs locally at the level of media items rather than domain concept and is only possible between items with explicit chronological dependencies.

Artequakt [6] The goal of the Artequakt project is to automatically generate biographies of artists from knowledge extracted from the web and maintained in a knowledge base. Each text fragment is annotated with the concepts from the domain ontology. This ontology can be extended dynamically with new instances and relationships. CIDOC is the main part of this domain ontology. In order to build a presentation Artequakt uses human authored templates of discourse structures. Templates are built for the biography genre. A template consists of a number of queries that can either retrieve the desired information from the database or construct sentences dynamically by retrieving specific facts from the ontology. A query is composed using domain classes and relationships between them. For example: ?artist died ?date_of_death in ?place_of_death. The overall structure of a template consists of several sub-structures to define the order in which concepts should be presented: *Sequence*, *Concept* and *Level Of Detail*. The basic *Sequence* structure defines an ordered list of queries that are instantiated from the database. *Concept* represents a set of queries rather than a sequence meaning that any of the queries can be executed in that point of the story building process. A *Level Of Detail* structure allows definition of the ordering preference in the query instantiation meaning that if

the highest numbered query cannot be used the next highest should be taken. A template can also contain additional contextual information allowing to adjust to different user characteristics (expert vs. novice).

Summary:

- The ordering rules are defined within a template but their interpretation occurs during the generation process. There is no explicit dependency between semantics of queries in a template and their arrangement within ordering sub-structures (*Sequence, Concept, Level Of Detail*).
- These ordering rules and query compositions are hard-coded, meaning that there is no rule external to a template that can guide the creation of a discourse structure.

DISC [5] uses the annotated multimedia repository of the Rijksmuseum and a domain ontology to create multimedia presentations on request. The aim of the approach is to build a multimedia presentation about a certain topic by traversing a semantic graph. The semantic graph consists of domain ontology of classes, instances and relations between them together with the media material related to those instances. There is a one-to-one relation between an instance in the domain ontology (e.g. Rembrandt) and a media item representing this instance (e.g. a self-portrait of Rembrandt). To create a discourse structure the system contains a set of rules explicitly described in RDF/S terms.

These rules define:
- what kind of genre can be applied to a certain main character (a biography for a person);
- what types of narrative units are relevant for a certain genre (a narrative unit describing personal life);
- what types of characters can appear in what narrative unit (a *Personal Life* unit talks about the main character (a person) but can also talk about a spouse of a person as a related character). Characters are mapped to domain classes (Person, Artist).
- what types of domain relationships are relevant for those characters. Inside a *Personal Life* narrative unit isMarried relation leads for example to a Spouse related character.

A set of narrative units defines a discourse structure for a presentation. This set is connected to a genre via a template. Such an architecture allows flexible development of a story inside a narrative unit since a number of related characters it describes depends on the information found in the semantic graph. For instance, the *Personal Life* narrative unit can talk about spouse and children of a main character. In the case of Rembrandt as the main character the semantic graph contains information about his wife Saskia and his son Titus. Thus, the presentation will be extended with the description of these two characters. In the case of Caravaggio information about his wife and children is absent. Therefore, the only information *Personal Life* unit will contain is information about Caravaggio himself. Further a related character can form a side-branch story by following domain relations that can be applied to this related character.

Summary:

- The specification of various discourse structures for the same genre and the same main character is not addressed. For example, a discourse structure for a biography genre contains *Personal Life* and *Career* narrative units. There are no rules that define that a biography can be built using a number of different discourse structures (one with the extended *Personal Life* narrative unit that elaborates on a large amount of the related characters and another one with the extended *Career* unit focusing more on person's achievements).
- There is no order defined for organizing narrative units into a discourse structure and for defining appearance of related characters inside a narrative unit.
- Besides, assuming a one-to-one relation between domain instances and media items, DISC cannot handle a real-world situation where multiple media items can be annotated with the same concept.

Samp*L*e [4] is a semi-automatic presentation generation environment that supports authors during a hypermedia presentation building process. The process is divided in four phases: topic identification, discourse structure building, media material collection and production of the final-form presentation. The aim of the Samp*L*e approach is to support authors during every phase of the process independent of the particular workflow. A workflow defines an order of proceeding from one phase of the process to another. Samp*L*e uses a semantic framework that combines existing thesauri in the art domain (such as AAT and ULAN translated in OWL) together with VRA schema for annotating images and Dublin Core as the top-level of the semantic framework. By integrating existing ontologies we try to enhance reuse of the semantic framework and annotated media items across different systems and approaches. We have also developed a discourse role ontology of concepts (e.g. introduction, description, elaboration, example) that determines possible roles media items can play in a particular discourse structure. Media material is annotated with concepts from domain and discourse ontologies. Discourse annotations of media items were initially integrated for facing the problem of dealing with multiple media items annotated with the same domain concept. Using discourse annotations allows to differentiate further between those media items and allows identifying their place within a discourse structure.

The first stage of Samp*L*e development aimed at supporting a workflow in which an author starts with defining a discourse structure and then has to collect media material to populate this structure.

For doing that the system had to have knowledge about:
- a genre that represents general properties of a discourse structure;
- various specific discourse structures that can be created for this genre[1];
- a way to extend existing discourse structures depending on the concrete main character selected;
- a mechanism to retrieve media items appropriate for the created discourse structure.

Initially discourse structures were built using human-authored templates. Thus, general principles of how a genre should be composed were not formalized. A template was built using domain classes to identify the content part of the discourse structure and discourse role concepts to specify appropriate types of media items. Classes in a template are arranged in a certain order to ensure coherence. The presence of domain classes in a template allowed to define an

[1] see `http://www.cwi.nl/~media/projects/CHIME/demos.html` for examples

overall discourse flow. The discourse structure extension mechanism allows to go from classes to domain instances by posing queries to the semantic framework. These queries retrieve instances applicable to the current main character. For example, a section in the discourse structure template *Members of a movement* for the main character De Stijl is extended with a number of subsections each talking about one of the De Stijl artists.

To define a mapping between a discourse structure and discourse role annotations each discourse structure was divided into Prologue, Main and Epilogue parts. Each of these parts were related to a number of discourse role annotations depending on the genre of a discourse structure.

Summary 1:

- Semantic relations between classes in a template were not explicitly defined. For example, a template for an essay about a Movement contained a sequence of domain classes. This sequence specified that we first would like to present a Movement, then talk about Principles and Artists. Here Movement, Principle, Artist are classes in the domain ontology. It was implicitly assumed that these principles and artists should be related to the concrete movement that is the topic of a specific presentation (e.g. Cubism). Thus, the generation process required additional rules for identifying whether different domain concepts are related.
- There was also a problem similar to one in the Artequakt approach: genre composition principles including ordering instructions were not defined with rules but were "hard-coded" in a template and interpreted during the generation process.

In the second stage of the Samp*L*e project we address the reverse workflow where an author first collects media material from the repository while browsing and then the system has to arrange this material into a coherent discourse structure. For supporting this workflow we wanted to reuse templates developed for the first workflow. For that we needed to provide an extension mechanism similar to the first workflow. This mechanism should replace domain classes in a template with instances. The difference in this case is that those instances are retrieved now from the concepts present in the selected media items annotations rather than from the semantic framework.

After extending a discourse structure with instances we could do a simple mapping between media item annotations and instances in the template. We found out that this process omits many related concepts from being included into a discourse structure (and thus prevents media items annotated with those concepts to appear in the presentation). This problem occurred due to the very simple manner in which templates were created. Basically the only information that a template contains is a sequence of domain classes that should appear in a discourse structure. Thus, there is no means to specify where in a discourse structure each particular instance should appear based on its relation to the main character. For instance, this is applicable to a case when we would like to specify that: if the main character is a movement and there are other related movements, they should be divided between two sections: *Preceding* and *Following* movements based on the chronological relations between them. For this rule we would have to identify first which relation identifies relevance of one movement to another (e.g. two movements from the same period are related) and then how based on this relations the movements have to be distributed within the discourse structure.

We also noticed that in the case when an artist is the main character and there are other related artists, a different type of domain relations (rather than a time-based one) should be applied to identify relevance between artists and a place for those artists in a discourse structure. To resolve this problem we created a number of rules which enabled inclusion of all related concepts in to a discourse structure.

Summary 2:

- The major part of system's knowledge was procedural and included into a generation model.
- Besides, continuing testing our rules on different sets of media items we were discovering new rules that should be included in the system. Each update causes the generation process to expand. In addition, existing rules in the generation process were not always re-used due to particular data structures this process operates upon. The selection of the most appropriate data structure was not possible at the beginning, since the complete system configuration and functionality was unknown. Thus, there was no efficient re-use of processes or knowledge even within one system.
- Absence of explicit genre composition principles also caused absence of explicit rules for managing various related characters. We discovered that in order to provide these rules the system would need a definition of what relationships between two domain concepts are meaningful for a certain discourse structure.

3. REQUIREMENTS FOR A COMMON DISCOURSE MODEL

Our current aim is to develop a discourse model that can cope with different types of discourse structure authoring approaches. This model should contain explicit discourse structure building principles to enable re-use within and across applications. We analyzed different components and prerequisites of the approaches and have come up with a set of requirements for a discourse model. In this section we will outline these requirements. The description will be based on the arguments presented in the previous section.

None of the systems described above uses an explicit representation of genre building principles. This caused a number of system components to be designed in a way that hinders flexibility and reusability:

- no explicit dependences between semantics of discourse structure elements and their order;
- ordering rules are defined and interpreted in the generation model;
- no possibility to specify alternative discourse structures for the same genre and the same main character;
- semantic relationships between classes are not explicitly defined in a template, and thus should be processed by the generation process;
- the generation process has to include different rules for managing various related characters;

As we can see the majority of these effects force the generation model to contain a large amount of knowledge for creating discourse structures. Since genre defines the global composition of a discourse structure, the first requirement for making discourse building knowledge explicit is:

R1: *Genre composition principles should be the first explicitly expressed within the model.*

Then all the levels of abstraction can be clearly defined: genre structure, discourse structure and the mapping between a discourse structure and media items. Genre modeling principles will guide discourse structure modeling. The questions about ordering of domain concepts within a discourse structure and identifying places for various related characters could be answered based on genre modeling principles and thus could be re-used for all similar discourse structures.

In order to handle multiple media items annotated with the same domain concepts we should

R2: *Provide means for distinguishing between these items depending on what genre and discourse structure they are used in.*

These means could be some sort of descriptions which then should be mapped into discourse structure components.

Based on our Samp*L*e and DISC experiences we realized that the way domain concepts are ordered inside a discourse structure depends on the genre and their relations to the main character. Thus,

R3 *A discourse model should be based on a structure that defines how to identify related characters for a particular discourse structure and how to place then within this structure.*

The analysis above still leaves a number of open questions:

- How exactly should genre modeling principles be encoded? It is the most difficult question of all since 1) there are different levels of abstraction in the model (domain classes, domain instances, annotated media items); 2) the rules have to address the instance level but not to be too specific; 3) ordering instructions should be applicable on the instance level but defined generally for a genre.
- What knowledge should be present in a genre definition and what knowledge should be present in a discourse structure?

4. CONCLUSIONS

In this paper we investigated different approaches for discourse structure authoring in a number of hypermedia presentation generation systems. We presented our motivation to work towards a discourse model, where discourse-related principles are explicitly encoded and can be re-used and exchanged. Based on our analysis of existing approaches we came up with the set of requirements which such discourse model should satisfy. Additionally, we highlighted a number of still unanswered questions. During the workshop we would like to discuss the requirements and their prerequisites with workshop participants. We might find out during the discussion that there are still some components or dependencies that we have not taken into account.

5. ACKNOWLEDGMENTS

This research is funded by the Dutch National NWO ToKeN2000 CHIME and I2RP projects. The authors would like to thank Frank Nack and Jacco van Ossenbruggen for the valuable comments and discussions during the development of this work.

6. REFERENCES

[1] L. Breure. Reuse of Content and Digital Genres. In H. van Oostendorp, L. Breure, and A. Dillon, editors, *Creation, Use and Deployment of Digital Information*, pages 27–53. Lawrence Erlbaum Associates, 2005.

[2] 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.

[3] L. M. Carter. Arguments in Hypertext: A Rhetorical Approach. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*, pages 87–91, San Antonio, Texas, USA, May 30 – June 3, 2000. ACM.

[4] K. Falkovych and F. Nack. Context Aware Guidance for Multimedia Authoring: Harmonizing Domain and Discourse Knowledge. *Multimedia Systems Journal*, 2006.

[5] 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.

[6] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal. Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. Presented at the Semantic Authoring, Annotation and Knowledge Markup (SAAKM) 2002 Workshop at the 15th European Conference on Artificial Intelligence (ECAI 2002), Lyon, France.

[7] C. Lagoze and H. V. de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. *JCDL2001*, 2001.

[8] Rijksmuseum Amsterdam. Rijksmuseum Amsterdam Website. http://www.rijksmuseum.nl.

[9] L. Rutledge, M. Alberink, R. Brussee, S. Pokraev, W. van Dieten, and M. Veenstra. Finding the Story — Broader Applicability of Semantics and Discourse for Hypermedia Generation. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pages 67–76, Nottingham, UK, August 23-27, 2003. ACM.

[10] 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.

[11] 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.