An XML/RST-based approach to multilingual document generation for the web

Guillermo BARRUTIETA
Mondragon Unibertsitatea
Loramendi, 4
Arrasate, Spain, 20500
gbarrutieta@eps.muni.es

Joseba ABAITUA
Universidad de Deusto
Avenida de las Universidades, 24
Bilbao, Spain, 48007
abaitua@fil.deusto.es

JosuKa DÍAZ
Universidad de Deusto
Avenida de las Universidades, 24
Bilbao, Spain, 48007
josuka@eside.deusto.es

Abstract: This paper shows how the framework of Rhetorical Structure Theory (RST) for discourse modelling can be expressed through XML annotations and then used to implement a natural language generation (NLG) system for the web. The system applies simplified RST schemes to the elaboration of a master document in XML from which content segments are chosen to suit the user's needs. The personalisation of the document is achieved through the application of a sequence of filtering levels of content selection based on the user aspects given as input.

1 Introduction

XML and its related standards (XSL, XSLT, DOM, SAX, etc.) are becoming a key technology in the processing and interchange of all types of documentation on the internet. In spite of this, XML's impact on the Natural Language Processing community is up to now moderate, and largely related to its capacity as standard for data representation, particularly for the annotation of language corpora. In this paper, we show how XML can play a much more central role in NLP through the description of a multilingual document generation system.

The validity of SGML/XML for natural language generation (NLG) has been proven before. Casillas et al. (1999, 2000) showed how DTDs derived from bilingual annotated corpora could be employed to generate new bilingual documents. Paired DTDs were used to construct document templates, which held the logical structure of the documents in each language and anticipated large parts of their content. Wilcock (2001) described XML-based tools to implement well-known approaches to NLG, such as pipeline architecture, templates, or tree-to-tree transformations. Here we discuss how the Rhetorical Structure Theory (RST) of Mann and Thompson (1988), which is a widely-used framework for NLG, can be expressed through XML annotations and then used to generate personalised documents in a learning environment for the web. The system is called the CourseViewGenerator (Barrutieta, 2001) and automatically produces learning objetcs that suit the student's needs at each particular stage of the learning process. We will explain the different parts of the system focusing on the components that more heavily rely on XML/XSL, such as the representation of RST in the form of XML elements and the content selection algorithm implemented through a sequence of XSL filters.

2 Gross-grained RST

The system starts by constructing a master document of the kind Hirst et al. (1997) proposed. This master document consists in a full-fledged text with references to all necessary multimedia elements (figures, tables, pictures, links, etc.) with all relevant information tagged in XML. The text is seen as raw data, and tags encapsulate these raw data as metadata reflecting the discourse structure of the text. This is represented using a simplified version of RST (Barrutieta et al., 2001). RST is simplified in the sense that the granularity of discourse segments does not transcend the boundaries of the sentence. Still, as any other standard RST discourse representation tree, this simplified version of RST contains a nucleus for each text paragraph, and one or several satellites linked by a discourse relation to the nucleus within the same paragraph. The nucleus is an absolutely essential segment of the text, as it carries the main message that the author wants to convey.

Satellites can be replaced or erased without changing the overall message and play an important supporting role for the nucleus.

```
<RST>
 <RST-S>
  <PREPARATION>
   <S>
    What is knowledge management?
   </S>
  </PREPARATION>
 </RST-S>
 <RST-N>
  <S>
   Knowledge, in a business context, is the organizational
   memory, which  people know collectively and individually
  </S>
  <S>
   Management is the judicious use of means to accomplish
   an end
  </S>
  <S>
   Knowledge management is the combination of those
    concepts, KM = knowledge + management
  </S>
 </RST-N>
</RST>
```

```
<RST>
 <RST-S>
  <PREPARATION>
   <S>
    ¿Qué es gestión del conocimiento?
   </S>
  </PREPARATION>
 </RST-S>
 <RST-N>
  <S>
   Conocimiento, en el contexto de los negocios, es la
   memoria de la organización, lo que la gente sabe colectiva
   e individualmente
  </S>
  <S>
   Gestión es el uso juicioso de recursos para alcanzar un fin
  </S>
  <S>
   Gestión del conocimiento es la combinación de esos dos
   conceptos, GC = gestión + conocimiento
  </S>
 </RST-N>
</RST>
```

```
<RST>
 <RST-S>
  <PREPARATION>
   <S>
    Zer da ezagutzaren kudeaketa?
   </S>
  </PREPARATION>
 </RST-S>
 <RST-N>
  <S>
   Kudeaketa, negozioetan, erakundearen memoria
   da, jendeak bakarka eta taldeka dakiena
  </S>
  <S>
   Kudeaketak erabideen erabilera zuzena du helburu
  </S>
  <S>
   Ezagutzaren kudeaketa bi kontzeptu hauen nahasketa da,
   EK = ezagutza + kudeaketa
  </S>
 </RST-N>
</RST>
```

Table 1: Gross-grained RST in XML

All the RST constituents (the nucleus together with the satellites) are represented in the form of XML elements. This notational variant adds very little novelty to RST but shows an effective way of bringing the theory closer to a practical application for the web.

```
<!ELEMENT SUBJECT (ADMIN,COURSE+ )>
<!ELEMENT ADMIN
(SUBJECTNAME,DEGREE,MOTIVATION,TIMEDISTRIBUT
ION,LANGUAGE,PROFESSORS,GOALS,
THEORETICALCONTENT,PRACTICALCONTENT,MATERIAL
,METHODOLOGY, EVALUATION,REFERENCES )>
<!ELEMENT SUBJECTNAME (#PCDATA)>
<!ELEMENT DEGREE (S+ )>
<!ELEMENT MOTIVATION (S+ )>
<!ELEMENT TIMEDISTRIBUTION
(CREDITS,HOURS,THEORY,EXERCISES,LAB,YEAR,SEM
ESTER,HOURSAWEEK )>
<!ELEMENT COURSE (INTRO,LESSON+,CONCLUSION
)>
<!ATTLIST COURSE
  LANG (ES|EN|EU)  #REQUIRED>
<!ELEMENT INTRO  (COURSE_TITLE,S+ )>
<!ELEMENT COURSE_TITLE  (S+ )>
<!ELEMENT LESSON  (TITLE , EXPLANATION+)>
<!ATTLIST LESSON
  DAY (1|2|3|4|5|6|7|8|9|10) #REQUIRED>
<!ELEMENT CONCLUSION (S+)>
<!ELEMENT TITLE  (S+ )>
<!ELEMENT EXPLANATION (RST+ )>
<!ELEMENT RST (RST-S|RST-N)*>
<!ELEMENT RST-N (S|
                 RST|
                 CONTRAST|
                 JOINT|
                 SEQUENCE|
                 LIST)*>
<!ELEMENT RST-S (EVIDENCE|
                 BACKGROUND|
                 ELABORATION|
                 ELABORATION-LINK|
                 ELABORATION-IMAGE|
                 PREPARATION|
                 ANTITHESIS|
                 CIRCUMSTANCE|
                 CONDITION|
                 ENABLEMENT|
                 EVALUATE|
                 INTERPRETATION|
                 JUSTIFY|
                 MOTIVATE|
                 NON-VOLITIONAL-CAUSE|
                 NON-VOLITIONAL-RESULT|
                 OTHERWISE|
                 PURPOSE|
                 RESTATEMENT|
                 SOLUTIONHOOD|
                 SUMMARY|
                 VOLITIONAL-CAUSE|
                 VOLITIONAL-RESULT|
                 EXAMPLE|
                 EXERCISE|
                 CONCESSION)*>
<!ELEMENT CONTRAST (S+)>
```

Table 2: DTD showing the structure of Learning Objects

It is very difficult to anticipate which relation-satellites will be needed to define a given discourse (Knott, 1995). Some of the most commonly used relations have been classified by Hovy & Maier (1997)., Here we will only mention those used in the master document taken as example. The notational transformation involves the rewriting of RST structures in the form of a Document Type Definition (DTD). The DTD is used to declare the names and properties of all those relations that will make up the logical structure of the document (or discourse, in the sense of RST). For example, EXPLANATION is a necessary part of the logical element LESSON, which in turn, together with INTRO and CONCLUSION, forms part of more abstract element defined as COURSE. An EXPLANATION consists of an obligatory nucleus and several optional elements, including BACKGROUND, ELABORATION, PURPOSE, and other similar relations.

A content selection algorithm that crucially depends on the user aspects given as input will carry out the generation of the final learning document, by selecting or discarding among all candidate relations.

3 User Aspects

User modeling is a key factor for successful document generation. This section describes the user aspects defined as the results of a survey among professionals and students (Table 3). These were asked to validate the properties of good documentation for training purposes, where document means readable and easy to understand.

User Aspect	**Frequency**
Knowledge about the subject matter before reading the document	34
Time available to read the document	29
Reason to read the document	27
Age	19
Education and education level	18
Language and nationality	17
Social, economic and cultural situation and level	12
Preference of text versus images	12
Preference towards access structures of text (indexes, …)	8
Number of readers and diversity	6
Preference towards rhetorical structures	5
Job	5
Interest in other subjects (other than the subject of the document)	5
Location when the document is read	5
Opinion about the subject (if any)	5
Preferences towards bibliographical references and links to other documents	5
Gender	5
Relation to the subject of the document	4
Personal situation (busy, tired, …)	1

Table 3: User Aspects

The content selection algorithm outlines the document structure based on the reader's profile defined through a set of multi-value parameters (Barrutieta et al, 2003). From these pondered user aspects, a user model was inferred. Table 4 illustrates a simplified version of the model.

Specific User Aspects	Discrete values
Subject	Language processors
Moment in time	Before the course/ Period 1/ Period 2/ … / After the course (review)
Languages	EN/ ES/ EU
General User Aspects	**Discrete values**
Level of expertise	Null/ Basic/ Medium/ High
Reason to read	To get an idea/ To get deep into it
Background	Not related to the subject/ Related to the subject
Opinion or motivation	Against/ Without an opinion or motivation/ In favour
Time available	A little bit of time/ Quite some time/ Enough time

Table 4: User model

4 Content selection

Content selection is a key factor of any successful NLG system. This process is normally seen as a goal-directed activity in which text segments are fit into the discourse structure of the text so as to convey a coherent communicative goal (Grosz and Sidner 1986). Content planning techniques, such as textual schemas (McKeown 1985) or plan operators (Moore and Paris 1993), have been successfully used as models of text generation. There are cases, though, in which these techniques may face some limitations, for example, when the structure of the discourse is difficult to anticipate (Mellish et al. 1998). Nevertheless, when a set of well-defined communicative goals exists, complex goals can be broken down into sequences of utterances and generation becomes an efficient "top-down" process (Marcu 1997).

In this section we show a content selection algorithm that works at a macro level with the parameters given by the user profile. The selected segments will make explicit the parameters that better reflect the user's profile within the model. In principle, nuclei will always be chosen (as they convey the main message of the text); satellites, however, will be selected depending on their relation to the nucleus and the user aspects that are activated at the time of generation.

The selection algorithm works in three consecutive phases: parallel selection, horizontal filtering and vertical filtering. Vertical filtering is the most important phase of the three as it is here that the parts of the discourse tree are selected or discarded.

4.1 CSA - Parallel selection - Phase 1

In the phase of parallel selection two of the three specific user aspects are taken into account: subject and languages. These aspects identify the relevant XML master document in the chosen language (as illustrated in figure 2.). There is one master document for each subject covered by the system, and these documents contain parallel aligned versions of the texts in each language (English, Spanish and Basque, in our case).

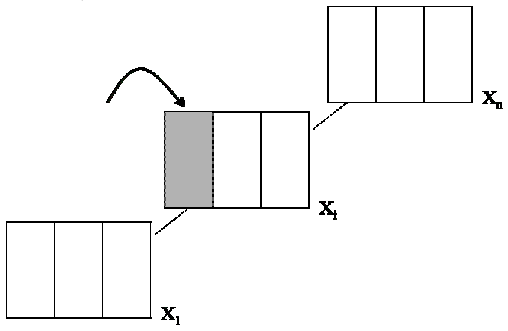

Figure 1: CSA – Parallel selection

As a result of this first filtering phase, the appropriate language division of the master document is selected. This text division is the input for subsequent filtering phases in which the particular segments of the document will be discriminated.

4.2 CSA - Horizontal filtering - Phase 2

The horizontal filtering phase concerns the third remaining user aspect that is moment in time, which is used to suit the generated text to the particular moment of the learning plan. This aspect cuts horizontally the parallel selection of the previous section.

The master document is structured in accordance with a set of course scheduling parameters. Each day and learning unit within the day is correlated with corresponding set of learning entities in the XML master document. In this way, the generated document can be targeted for learning unit 1 of day 1, or any other day or unit. The XML master file also contains some informative elements that the reader may need to know even before the course starts or after it has finished. These will be generated also as a result of some specific user aspects that are activated. Figure 3 shows a graphical representation of horizontal filtering.

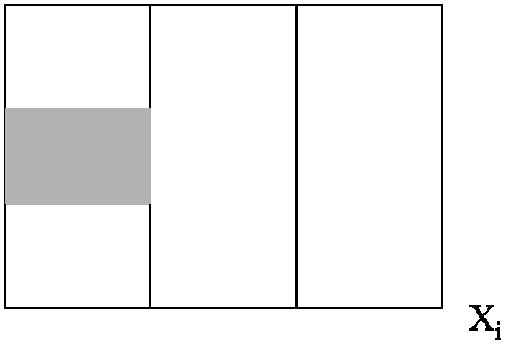

Figure 2: CSA – Horizontal filtering

4.3 CSA - Vertical filtering - Phase 3

The final phase of vertical filtering comprises the five user aspects of level expertise, reason to read, professional background, opinion or motivation and time available. These five aspects will be relevant to discriminate those parts of the discourse tree which have been previously selected and filtered.

Nuclei will be always maintained because they are, by definition, irreplaceable segments of the text and convey the main message. Satellites are segments of the text that will be subject to the algorithm's process of selection. Figure 4. shows graphically this filtering phase.

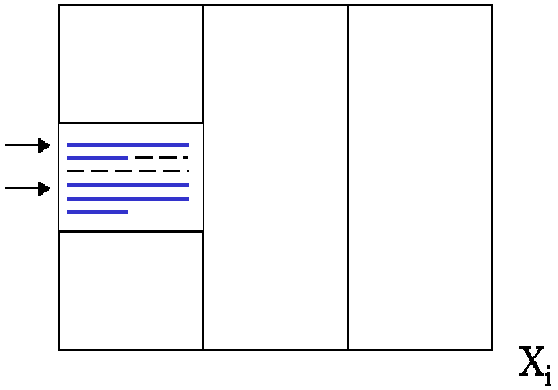

Figure 3: Vertical filtering

The set of discrimination rules applied in this first version of the content selection algorithm is described in Barrutieta et al. (2002). These rules apply in subsequent checking levels of filtering, and therefore have a cascading effect.

Cascading filters apply to the relation-satellites that are still active after the previous phases in the generation process. When a vertical filter 3 tries to get rid of a relation-satellite already abandoned at a previous phase (2 or 1), there will be nothing to act upon, but this circumstance will produce no consequence, since the CSA continues the filtering process on the remaining text. Thus, the order in which the vertical filters are applied is not relevant.

After the filtering process has been successfully completed, there is still a final presentation task. A good presentation is, in our opinion, one that will provide the student with an optimal version of the document to read, understand and fruitfully assimilate its content.

5 Implementation

The javascript code manages the user aspects (one of the inputs of the algorithm) and the application of the casdading filters (the CSA). Depending on the user aspects given by the user, the variables sXSL1 to sXSL5 take the value of the filter to be applied for each user aspect (level of expertise, reason to read, background, opinion or motivation and time available). The sResult variable contains the XML file whose content will be varying after each filter is applied. Table 3 shows the code that executes a filter.

```
objData.loadXML(sResult);
objStyle.load(sXSL1);
sResult=objData.transformNode(
objStyle);
```

Table 5: Javascript implementation

XSL filters pass on (or not) one element to the following vertical filter depending on the rules described before. Table 4 shows how this is done with the relation-satellite BACKGROUND.

```
<xsl:template
     match="BACKGROUND">
  <xsl:copy>
    <xsl:apply-templates/>
  </xsl:copy>
</xsl:template>
```

Table 6: XSL implementation

6 Conclusions

One of the features that is worth considering is the scalability of the filtering mechanism. We anticipate two types of expansions to the system: (1) Increasing the size of the corpus, including more subjects and master documents, and (2) augmenting the user model by adding user aspects or by adding more parameters to the existing user aspects.

The first type of expansion will not require any alteration of the CSA as long as the added document tokens conform to the existing DTD and our RST model. In order to increase the size of the corpus, it will be necessary to annotate XML discourse-tree metadata manually. This is a complex and time-consuming task (as has been noted by Carlson and Marcu, 2001). Future research activities should focus on helping automate the annotation process, for example using cue phrases à la Knott (Knott 1995; Alonso and Castellón, 2001).

The second type of expansion requires only the elaboration of additional XSL filters. Adding new values to existing user aspects requires only the modification of the corresponding XSL filter. Any of these last two operations can be incorporated easily. Therefore, adding a new user aspect or a new discrete value does not increase in any substantial way the complexity of the system. An on-line version of the system can be accessed at: http://www.muni.es/cvg

Acknowledgements

This research was partly supported by the Basque Government (XML-Bi, PI1999-72 project).

References

Alonso, L. and Castellón, I. (2001) Towards a delimitation of discursive segments for Natural Language Processing applications. First International Workshop on Semantics, Pragmatics and Rhetoric. Donostia (Spain), pp. 45-52.

Barrutieta, G. (2001) Generador inteligente de documentos de formación. Virtual Educa 2001, Madrid (Spain), pp. 256-261.

Barrutieta, G., Abaitua, J. and Díaz, J. (2001) Gross-grained RST through XML metadata for multilingual document generation. MT Summit VIII. Santiago de Compostela (Spain), pp. 39-42.

Barrutieta, G., Abaitua, J. and Díaz, J. (2001). Cascading XSL filters for content selection in multilingual document generation. Second Workshop on NLP and XML, Taipei (Taiwan)

Barrutieta, G., Abaitua, J. and Díaz, J. (2003) User modelling and content selection for multilingual document generation. VII Simposio Internacional de Comunicación Social, Santiago de Cuba.

Carlson, L.and Marcu, D. (2001) Discourse tagging manual. Technical report ISI-TR-545. ISI Marina del Rey (USA).

Casillas, A., Abaitua, J. and Martínez, R. (1999) Extracción y aprovechamiento de DTDs emparejadas en corpus paralelos. Procesamiento del Lenguaje Natural, 25:33-41.

Casillas, A., Abaitua, J. and Martínez, R. (2000) DTD-driven Bilingual Document Generation. International Natural Language Generation Conference. Mitzpe Ramon, Israel, 32-38.

Grosz, B. and Sidner, C. (1986) Attention, intentions and the structure of discourse", Computational Linguistics, 12:175-204.

Hirst, G., DiMarco, C., Hovy E. & Parsons K. (1997) Authoring and Generating Heath-Education Documents That Are Tailored to the Needs of the Individual Patient. Proceedings of the Sixth International Conference. UM97. Vienna (NY-USA), pp. 107-118.

Hovy, E. & Maier, E. (1997) Parsimonious or profligate: how many and which discourse structure relations?

Knott, A. (1995) A Data-Driven Methodology for Motivating a Set of Coherence Relations, Ph.D. thesis, University of Edinburgh, Edinburgh (UK).

Mann, W.C., and Thompson, S.A. (1988) Rhetorical Structure Theory: A theory of text organization. Tech. Rep. RS-87-190. Information Sciences Institute. Los Angeles, CA.

Marcu, D. (1997) From local to global coherence: a bottom-up approach to text planning, in Proceedings of AAAI-97, American Association for Artificial Intelligence, pp.629-635.

McKeown, K. (1985) Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text, Cambridge University Press.

Mellish, C., M. O'Donnell, J. Oberlander and A. Knott (1998) An architecture for opportunistic text generation. Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada, pp. 28-37.

Moore, J. and Paris, C. (1993) Planning texts for advisory dialogues: capturing intentional and rhetorical information, Computational Linguistics, 19.

Reiter, E. and Dale, R. (2000) Building applied natural language generation systems. Cambridge University Press (UK).

Wilcock, G. (2001) Pipelines, Templatges and Transformations: XML for Natural Language Generation. First Workshop on NLP and XML, Tokyo (Japan).