Gross-grained RST through XML Metadata for Multilingual Document Generation

Guillermo Barrutieta

Departmento de Informática
Mondragon Unibertsitatea
Loramendi, 4
20500 Arrasate (Spain)
gbarrutieta@eps.muni.es

Joseba Abaitua

Filosofía y Letras
Universidad de Deusto
Avda. de las Universidades, 24
48007 Bilbao (Spain)
abaitua@fil.deusto.es

JosuKa Díaz

ESIDE
Universidad de Deusto
Avda. de las Universidades, 24
48007 Bilbao (Spain)
josuka@eside.deusto.es

Abstract

We present an RST-based discourse annotation proposal used in the construction of a trial multilingual XML-tagged corpus of teaching material in Basque, English and Spanish. The corpus feeds an experimental multilingual document generation system for the web. The main contributions of this paper are an implementation of RST through XML metadata and the adoption of gross-grained RST to avoid non-isomorphism in multilingual corpora.

Keywords

Corpus, RST, XML

Background

The natural language generation community has conducted research in many aspects of multilingual generation systems. One of these aspects is the construction of multilingual corpora as the source to generation (Jurafsky et al. 2000).

In this paper we present a novel attempt to model a series of teaching units into an XML-tagged multilingual corpus for its use in a document generation system. The document generation system leans on a multilingual master document (Hirst et al., 1997) that contains text snippets that are combined in an adequate way to satisfy the needs of the users (or readers in this context) dynamically depending on the moment, level of expertise, time available and some other user aspects.

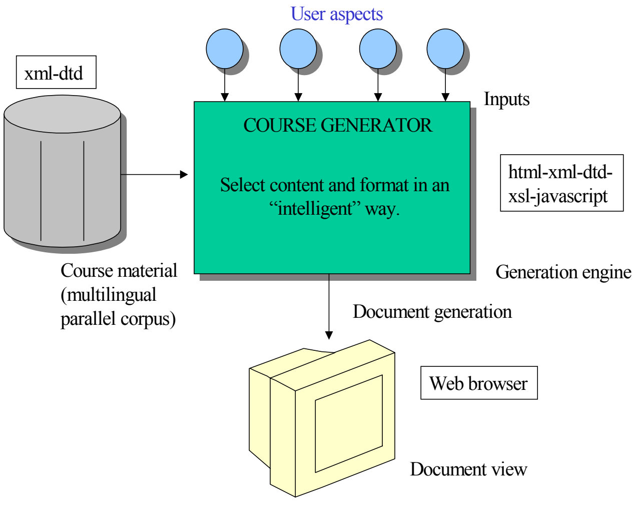
User aspects
xml-dtd
Inputs
COURSE GENERATOR
Select content and format in an "intelligent" way.
html-xml-dtd-xsl-javascript
Course material (multilingual parallel corpus)
Generation engine
Document generation
Web browser
Document view

Figure 1: General schema of the multilingual document generation system

The creation of the multilingual master document met a number questions that are listed below:

- What are the parts of a technical-educational document?
- What is the optimum segmentation level? Is it the clause, or the discourse unit?
- What are the relations among those parts or segments?
- In the domain of technical education, how can we formalize information at discourse level?
- How can this information be stored formally and in a reusable way for the purpose of document generation?

Rhetorical Structure Theory

Most of these questions found an adequate answer within the Rhetorical Structure Theory (RST) of Mann and Thompson (1988). RST granted a sound theoretical base and some helpful points of reference. According to RST, a text may be divided into minimal text spans that have an unambiguous discourse function (nuclei and satellites) and relations between these text spans can be established.

Documents in our corpus were segmented into minimal text spans that have an unambiguous discourse function; that is, nuclei which hold the core message and satellites that support the core message of the text. Nuclei retain the essential content to understand the text and satellites can be replaced or erased without changing the general meaning . RST caters also with a set of possible relations between nuclei and satellites. There is a wide-range of relations that link nuclei and satellites such as elaboration, background, justification, evidence, concession and so on. After manually annotating the discourse (Marcu 1999b) a tree is obtained. A discourse tree looks like this.

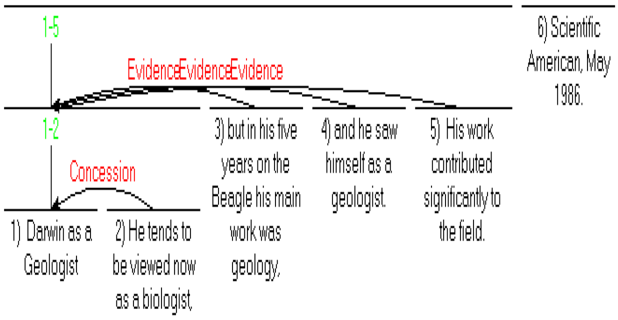
1-5
EvidenceEvidenceEvidence
6) Scientific American, May 1986.
1-2
Concession
3) but in his five years on the Beagle his main work was geology.
4) and he saw himself as a geologist.
5) His work contributed significantly to the field.
1) Darwin as a Geologist
2) He tends to be viewed now as a biologist,

Figure 2: RST discourse tree

RST through XML metadata

The technology of eXtensible Markup Language (XML) (Bray et al., 1998) provides a web standard and metalanguage to model mark-up that look like HTML tags. There is one major difference between XML tags and HTML tags, though: XML tags are created by the developer and unveil control information, that is, metadata, about the type of data that tags encapsulate. In our case, tags encapsulate parts of the text and unveil discourse level information for the document generation system.

The XML file that contains RST data and metadata looks like the example in figure 3. The DTD file that specifies the structure of the XML file is in figure 4.

Gross-grained RST through XML metadata

With RST we resolved the first set of problems. But RST also raises questions (Marcu et al., 2000) (Marcu et al., 1999a) about the non-isomorphism among the RST discourse-tree in multilingual corpora. We then faced the problem of how to address the likely isomorphism of a multilingual corpus. In other words, the discourse-tree for one language may be different from the needed tree of another language, and this lack of isomorphism seriously hinders the generation phase (Marcu et al. , 2000).

Marcu et al. propose (2000) a possible way of solving this problem: "For the purpose of multilingual text planning, one can, hence, assume that a language-independent text planner derives first a language-independent rhetorical structure and then linearizes it, i.e. transforms it to make it language specific".

```
.
.
.
<EXPLANATION>
 <RST>
  <RST-N>
    <RST>
     <RST-N>
       <S>
       Darwin as a geologist
       </S>
     </RST-N>
     <RST-S>
       <CONCESSION>
         <S>
         He tends to be viewed now as a biologist,
         </S>
       </CONCESSION>
     <RST-S>
    <RST>
  </RST-N>
  <RST-S>
    <EVIDENCE>
      <S>
      but in his five years on the Beagle his main work was
geology
      </S>
    </EVIDENCE>
    <EVIDENCE>
      <S>
      and he saw himself as a geologist.
      </S>
    </EVIDENCE>
    <EVIDENCE>
      <S>
      His work contributed significantly to the field.
      </S>
    </EVIDENCE>
  </RST-S>
 </RST>
</EXPLANATION>
.
.
.
```

Figure 3: RST discourse tree in XML

```
.
.
.
<!ELEMENT EXPLANATION (RST+ )>
<!ELEMENT RST (RST-S|RST-N)*>
<!ELEMENT RST-N (S|RST)*>
<!ELEMENT RST-S (
                EVIDENCE|
                CONCESSION)*>
<!ELEMENT EVIDENCE (S+)>
<!ELEMENT CONCESSION (S+)>
<!ELEMENT S (#PCDATA)>
.
.
.
```

Figure 4: DTD file that specifies the structure of the previous XML file

This language-independent rhetorical structure is what we capture through our gross-grained RST. Gross-grained

RST largely resembles RST (or full RST) in the sense that nuclei, satellites and the same set of relations are maintained. The difference is that text-spans dismiss the level of the smaller segments in the text, and range only over whole sentences, sets of sentences, full paragraphs or even entire documents, with the sole condition that they fulfil a communicative goal. These communicative goals include the act of preparing the reader for an incoming nucleus, an example or any evidence that supports the core message expressed by the nucleus, and other similar acts.

The corpus

This approach has been applied to a corpus with texts in three languages, English, Spanish and Basque. Gross-grained segmentation made it possible to maintain the isomorphic disposition of the corpus, despite its multilingual nature. A sample discourse-tree in XML can be seen in Figure 5.

Our document representations look very similar to standard RST. They are identical in terms of nuclei, satellites and relations, with the only exception of the granularity of the segmentation. Our segments are tuned to spelling conventions (period, colon, semicolon, full stop, etc.), which roughly reflect the structure of discourse. At this suprasentential level, it is not difficult to preserve parallelism among segments in different languages.

The generation algorithm

The generation system selects different parts of the text depending on the tag that formally stores discourse level information. The selection or discrimination of particular text segments is crucially determined by this discourse level information.

The set of user aspects that are requested for the selection or discrimination of a given text snippet is something that is now being developed.

Evaluation and future work

One important aspect that needs to be implemented is an evaluation methodology that can help adequately judge the validity of documents generated by the system.

We are also planning to increase the size of the corpus to demonstrate empirically whether the isomorphism of our gross-grained RST holds with a larger corpora.

```
<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>
```

Figure 5: Multilingual gross grained RST discourse tree in XML

Acknowledgements

We want to express our gratitude to Dr. Eduard Hovy and Dr. Daniel Marcu from the Information Sciences Institute at the USC for their invaluable contribution introducing us to the Rhetorical Structure Theory and for their advice and support during the summer 2000.

This research was partly supported by the Basque Government (XTRA-Bi OD-00UD05 and XML-Bi, PI1999-72 projects)

Bibliographical References

Bray, T. et al. (eds) (1998) Extensible Markup Language (XML) 1.0. World Wide Web Consortium <http://www.w3.org/TR/REC-xml>

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.

Jurafsky, D., Martin, J.H. (2000) Speech and Language Processing. Prentice Hall. Upper Saddle River, NJ.

Mann, W.C., Thompson, S.A. (1987) Rhetorical Structure Theory: A theory of text organization. Tech. Rep. RS-87-190. Information Sciences Institute. Los Angeles, CA.

Marcu, D. (1999b) Instructions for Manually Annotating the Discourse Structures of Texts. ISI-USC. Los Angeles, CA.

Marcu, D., Carlson L., Watanabe, M. (2000). An Empirical Study in Multilingual Natural Language Generation: What Should A Text Planner Do? The 1st International Conference on Natural Language Generation INLG'2000, Mitzpe Ramon, Israel.

Marcu, D., Romera, M., Amorrortu, E. (1999a). Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. The Workshop on Levels of Representation in Discourse, pages 71-78. Edinburgh, Scotland.