Structure-Preserving Pipelines for Digital Libraries

{massimo.poesio, eduard.barbu, egon.stemle}@unitn.it, cgirardi@fbk.eu

Summary

Existing HLT pipelines assume the input is pure text or, at most, HTML and either ignore (logical) document structure or remove it. But identifying the structure of documents is essential in digital library and other types of applications, and it is relatively straightforward to extend existing pipelines to achieve ones in which the structure of a document is preserved.

Problem

Freely available off-the-shelf (HLT) pipelines like LingPipe, OpenNLP, GATE, and TextPro support a variety of document formats as input. Still, actual processing rarely takes advantage of structural information.

When processing large documents, section or chapter boundaries are an important segmentation to use, and when working with the type of data typically found in digital libraries or historical archives, such as whole books, exhibition catalogues, scientific articles, contracts *document structure awareness* is vital.

Three types of problems with 'ignorant' HLT pipelines:

- Techniques for extracting content from plain text do not work on, e.g. bibliographic references, or lists
- Removing non-plain text parts of a document strips useful information, e.g. keywords, bibliographic references
- Parts that can be considered to contain plain text, e.g. titles, can be processed more appropriately when knowing about this being non-paragraph text

Logical Structure of Documents

Documents have at least two types of structure:

- Geometrical, or layout, structure, refers to the structuring of a document according to its visual appearance, its graphical representation (pages, columns)
- Logical structure refers to the content's organization to fulfil an intended overall communicative purpose (title, author list, chapter, section, bibliography)

Digital Libraries

HLT techniques in a digital library setting can be used for indexing documents in the library for search, or to classifying them to automatically extracting metadata. They are being incorporated in document management platforms and are used to support the librarians.

Document structure aware pipelines can suggest the most important keywords, find the text to be indexed or even summarized, and produce citations lists, possibly to be compared with the digital library's list of citations to decide whether to add them.

Intra- and Inter-Pipeline File Exchange Formats

HLT pipelines often eliminate mark-up during initial pre-processing. However, in contexts where different parts of the output of the pipeline need to be processed in different ways this is inappropriate. Still, instead of developing a new pipeline from scratch our goal has been to reuse existing pipelines and augment them with *document structure awareness*.

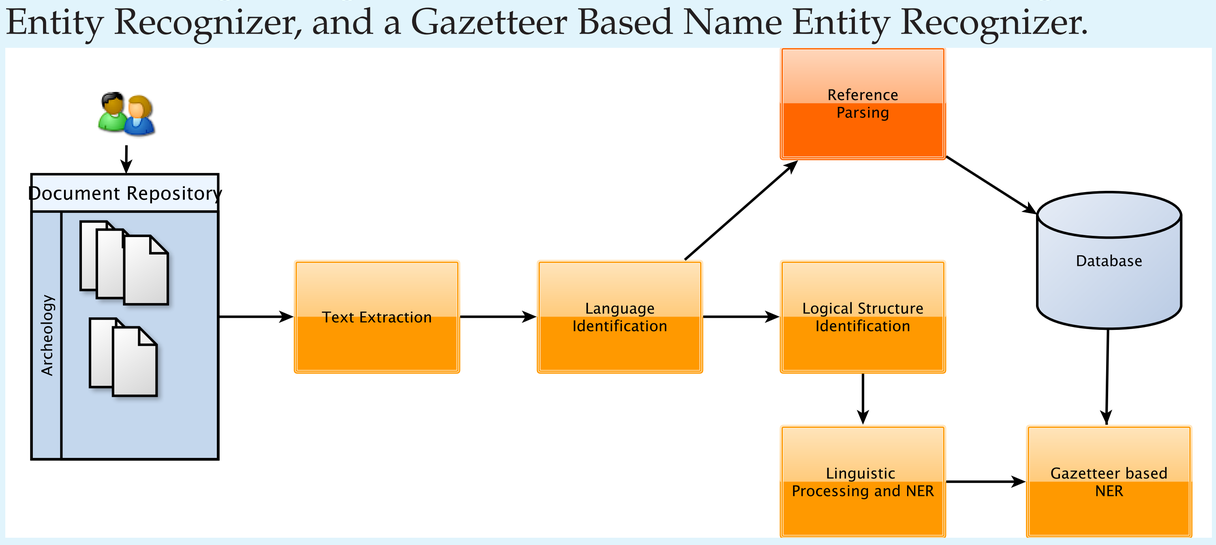
HLT pipelines mainly use tabular format, or inline or standoff XML format to pass on information from one processing stage to the next. To this end, ideally, both formats should be supported to facilitate intra-pipeline exchange.

```
# FILE: 11
# PART: id1
# SECTION: title
# FIELDS: token    tokenstart  sentence  pos    lemma         entity   nerType
Spondylus          0           -         SPN    Spondylus     O        B-SITE
gaederopus         10          -         YF     gaederopus    O        O
,                  20          -         XPW    ,             O        O
gioiello           22          -         SS     gioiello      O        O
dell'              31          -         E      dell'         O        O
Europa             36          -         SPN    europa        B-GPE    B-SITE
preistorica        43          -         AS     preistorico   O        O
.                  55          <eos>     XPS    full_stop     O        O
# FILE: 11
# PART: id2
# SECTION: author
# FIELDS: token    tokenstart  sentence  pos    lemma         entity   nerType
MARIA              0           -         SPN    Maria         B-PER    O
A                  6           -         E      A             I-PER    O
BORRELLO           8           -         SPN    Borrello      I-PER    O
&                  17          -         XPO    &             O        O
.                  19          <eos>     XPS    full_stop     O        O
```

```
<cesCorpus xsi:schemaLocation="http://www.xml-ces.org/schema a ➥
    http://www.cs.vassar.edu/XCES/schema/xcesDoc.xsd">
<xcesHeader xlink:href="Header11.xml"/>
   <cesDoc version="1.0">
      <text>
         <body>
         <div type="section" xml:lang="it">
            <p id="p1" type="title">
                <s id="p1s1"><name key="SITE1" type="site">Spondylus</name> gaederopus, ➥
                    a gioiello dell'<name key="GPE1" type="location">Europa</name> a ➥
                    preistorica.</s>
            </p>
            <p id="p2" type="author">
                <s id="p2s1"><name key="PER1" type="person">MARIA A BORRELLO</name>&.</s>
            </p>
         </div>
         </body>
      </text>
   </cesDoc>
</cesCorpus>
```

Exemplified: A Pipeline for Processing Archaeological Articles

For processing research articles in Archaeology our pipeline integrates three main modules: one for recovering the logical structure of documents, one multi-lingual POS tagger and general Name Entity Recognizer, and a Gazetteer Based Name Entity Recognizer.



For extracting the logical structure of documents we use ParsCit. The particularities of the archaeology repository, however, require specifically trained CRF models. To this end, 55 documents (35 Italian, 20 English) have been annotated.

Further Applications

Document structure aware pipelines can be used for a variety of text-analysis tasks: analysis of blogs, microblogs, community QA sites, forums. Enhance search tasks by performing automatic query refinements by analysing, e.g. HTML documents' mark-up. Preserving structured information, e.g. from Wikipedia articles, like info boxes, categorization and geo information, links to other articles, to other wiki projects, and to external Web pages.