Hyperlinks analysis in Multilingual Web Applications

Filippo Ricca
ITC-irst
Centro per la Ricerca Scientifica e Tecnologica
ricca@itc.it

Multilingual Web applications

- To be accessible to a larger audience, Web applications are often required to be multilingual.
- A multilingual Web application is an application where the information contained in the Web pages is supplied in more than one language (e.g., Italian, English, Chinese, etc.).
- An example of it is the Gateway to the European Union (http://europa.eu.int/) where information is supplied in 20 different languages.
Quality of Web applications

- Quality of Web applications is very important.
- In general the quality delivered is often poor and this is particularly true for legacy multilingual Web applications.
- In multilingual Web sites information and structure should be consistent across languages.

In absence of CMS, development of multilingual sites is based on personal (non-standard) choices, and alignment during maintenance is achieved manually. This is particularly true for legacy Web applications.

Multilingual Web inconsistencies

<table>
<thead>
<tr>
<th>Absence of pages in some languages (missing translations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differences in the page structure in different languages</td>
</tr>
<tr>
<td>Missing information</td>
</tr>
<tr>
<td>Parts of the page not translated</td>
</tr>
<tr>
<td>Different internal or external hyperlinks.</td>
</tr>
<tr>
<td>Incorrect cross-language hyperlinks.</td>
</tr>
<tr>
<td>Incomplete change propagation across languages: pages not up-to-date.</td>
</tr>
</tbody>
</table>
**Missing translation**

Official site of the National Electro-technical Institute "Galileo Ferraris":

*Pagina4.html* is available only in Italian. *History.html, present.html* and *staff.html* are only in English.

**Unintend language switch**

Official site of the National Electro-technical Institute "Galileo Ferraris":

If *Pagina4.html* is reached from an Italian page and the standard time is accessed the English version of the page is showed.
Alignment toolkit

To cope with Web inconsistencies we devised a number of algorithms implemented and tested in a toolkit.

Language identification

Please insert: name, surname

STOP - Please insert: nome, cognome

ACK ESC: up
DEL INS: down

<table>
<thead>
<tr>
<th>Language</th>
<th>English dict.</th>
<th>Italian dict.</th>
<th>err/words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0/4</td>
<td>4/4</td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td>2/5</td>
<td>2/5</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>4/6</td>
<td>6/6</td>
<td></td>
</tr>
</tbody>
</table>

Common mistakes:

- Added a character
- Changed a character
- Added a word
- Changed a word
- Reordered words
- Changed part of a word
- Changed number of characters

- Added a character
- Changed a character
- Added a word
- Changed a word
- Reordered words
- Changed part of a word
- Changed number of characters
- Added a character
- Changed a character
- Added a word
- Changed a word
- Reordered words
- Changed part of a word
- Changed number of characters
Page matching: structure

\[
\text{(A, B, C, /C, D, /D, /B, B, /B, /A)} \quad \text{(A, B, C, /C, /B, B, /B, /A)}
\]
\[\text{del(D), del(/D)} \quad \text{dist} = 2\]

Edit distance: content

\[
\text{(A, B, C, /C, D, /D, /B, B, /B, /A)} \quad \text{(A, B, C, /C, /B, B, /B, /A)}
\]
\[\text{del(D), del(/D), del(text2), ins(transl(text6))} \quad \text{dist} = 4\]
Page alignment

- Output of page matching: pairs of corresponding pages in different languages.
- *Edit script:* edit operations to be performed to align corresponding pages (including text translations).

The Web site developer can update pages making them cross-language consistent

Internal hyperlink inconsistency

Corresponding pages *p-en*, *q-en* should also be connected by a hyperlink.
Cross-language hyperlinks: external hyperlink inconsistencies (type 1)

A hyperlink from p-en to p-it should also exist.

Cross-language hyperlinks: external hyperlink inconsistencies (type 2)

A hyperlink from p-en to q-it should also exist.
FOR EACH $a_{lang1} \rightarrow b_{lang1} \in \text{PORTION1}$ {
    $a_{lang2} = \text{pageMatching}(a_{lang1})$;
    $b_{lang2} = \text{pageMatching}(b_{lang1})$;
    IF $a_{lang2} \rightarrow b_{lang2} \notin \text{PORTION2}$
    Int_Set = Int_Set $\cup$ \{$a_{lang2} \rightarrow b_{lang2}$\}
}
FOR EACH $a_{lang2} \rightarrow b_{lang2} \in \text{PORTION2}$ {
    $a_{lang1} = \text{pageMatching}(a_{lang2})$;
    $b_{lang1} = \text{pageMatching}(b_{lang2})$;
    IF $a_{lang1} \rightarrow b_{lang1} \notin \text{PORTION1}$
    Int_Set = Int_Set $\cup$ \{$a_{lang1} \rightarrow b_{lang1}$\}
}
FOR EACH $a_{lang1} \rightarrow b_{lang2} \in \text{ALL}$ {
    $a_{lang2} = \text{pageMatching}(a_{lang1})$;
    $b_{lang1} = \text{pageMatching}(b_{lang2})$;
    IF $b_{lang2} == a_{lang2}$ { // Type 1
        IF $b_{lang2} \rightarrow a_{lang1} \notin \text{ALL}$
        Ext1_Set = Ext1_Set $\cup$ \{$b_{lang2} \rightarrow a_{lang1}$\};
    } ELSE { // Type 2
        IF $a_{lang2} \rightarrow b_{lang1} \notin \text{ALL}$
        Ext2_Set = Ext2_Set $\cup$ \{$a_{lang2} \rightarrow b_{lang1}$\};
    }
}

Page merging

HTML $\rightarrow$ Tidy $\rightarrow$ XHTML $\rightarrow$ PageMerger $\rightarrow$ MLHTML

p-it.html $\rightarrow$ p-it.xhtml $\rightarrow$ p-ml
p-en.html $\rightarrow$ p-en.xhtml

where p-it.html and p-en.html are aligned.
MLHTML: Multilingual XHTML

- MLHTML is an XML representation of multilingual Web pages
- Information in all languages is inserted into a single file
- Document format and structure are shared among languages
- The multilingual content is inserted inside adjacent lines

```xml
<title>
<ml lang="en"> Publication List </ml>
<ml lang="it"> Lista delle pubblicazioni </ml>
</title>
```

Experimental Results

<table>
<thead>
<tr>
<th>Web sites</th>
<th>Pages</th>
<th>LOC</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.sitaserre.com">www.sitaserre.com</a></td>
<td>18</td>
<td>2732</td>
<td>252</td>
</tr>
<tr>
<td><a href="http://www.albergghiero-longarone.it">www.albergghiero-longarone.it</a></td>
<td>25</td>
<td>8028</td>
<td>182</td>
</tr>
<tr>
<td><a href="http://www.mediaeletra.com">www.mediaeletra.com</a></td>
<td>37</td>
<td>29954</td>
<td>2144</td>
</tr>
<tr>
<td><a href="http://www.artifer.com">www.artifer.com</a></td>
<td>54</td>
<td>10827</td>
<td>1003</td>
</tr>
<tr>
<td><a href="http://www.toscana-toskana.de">www.toscana-toskana.de</a></td>
<td>68</td>
<td>13734</td>
<td>1240</td>
</tr>
<tr>
<td><a href="http://www.trevirtu.com">www.trevirtu.com</a></td>
<td>81</td>
<td>17522</td>
<td>870</td>
</tr>
<tr>
<td><a href="http://www.gigirrosso.com">www.gigirrosso.com</a></td>
<td>84</td>
<td>11195</td>
<td>1689</td>
</tr>
<tr>
<td><a href="http://www.agenziarossi.com">www.agenziarossi.com</a></td>
<td>125</td>
<td>55587</td>
<td>558</td>
</tr>
<tr>
<td><a href="http://www.viniesaporidipuglia.com">www.viniesaporidipuglia.com</a></td>
<td>210</td>
<td>28036</td>
<td>2418</td>
</tr>
<tr>
<td><a href="http://www.olio-oliva.it">www.olio-oliva.it</a></td>
<td>275</td>
<td>25425</td>
<td>809</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>977</strong></td>
<td><strong>203040</strong></td>
<td><strong>11165</strong></td>
</tr>
</tbody>
</table>

- 10 Web applications chosen among Italy-based firms and educational institutions has been downloaded by means of the **ReWeb** tool.
Hyperlinks inconsistencies revealed

<table>
<thead>
<tr>
<th>Web sites</th>
<th>Int</th>
<th>Ext 1</th>
<th>Ext 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.sitaserre.com">www.sitaserre.com</a></td>
<td>4/36</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td><a href="http://www.alberghiero-longarone.it">www.alberghiero-longarone.it</a></td>
<td>10/36</td>
<td>0/8</td>
<td>7/21</td>
</tr>
<tr>
<td><a href="http://www.medialettra.com">www.medialettra.com</a></td>
<td>23/648</td>
<td>1/3</td>
<td>1/27</td>
</tr>
<tr>
<td><a href="http://www.artifer.com">www.artifer.com</a></td>
<td>0/72</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td><a href="http://www.toscana-toskana.de">www.toscana-toskana.de</a></td>
<td>9/90</td>
<td>2/8</td>
<td>6/28</td>
</tr>
<tr>
<td><a href="http://www.trevirtu.com">www.trevirtu.com</a></td>
<td>9/377</td>
<td>1/51</td>
<td>9/9</td>
</tr>
<tr>
<td><a href="http://www.gigirosso.com">www.gigirosso.com</a></td>
<td>0/342</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td><a href="http://www.agenziarossi.com">www.agenziarossi.com</a></td>
<td>16/301</td>
<td>0/0</td>
<td>0/1</td>
</tr>
<tr>
<td><a href="http://www.viniesaporidipuglia.com">www.viniesaporidipuglia.com</a></td>
<td>24/2192</td>
<td>0/2</td>
<td>7/21</td>
</tr>
<tr>
<td><a href="http://www.olio-oliva.it">www.olio-oliva.it</a></td>
<td>18/661</td>
<td>0/2</td>
<td>12/13</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>113/4755</td>
<td>4/74</td>
<td>42/120</td>
</tr>
</tbody>
</table>

www.sitaserre.com

Four outgoing links of form-eng.html have no corresponding links outgoing from form.html.

Missing hyperlinks
( Int, form.html ----> ita/index.html, Italian)
( Int, form.html ----> ita/qualita.html, Italian)
( Int, form.html ----> ita/span.html, Italian)
( Int, form.html ----> ita/altri.html, Italian)
Conclusions and future work

- We have explained in detail a module of the Alignment Toolkit: the Hyperlinks analyzer.
- The Alignment Toolkit helps the Web maintainer to re-structure an existing Web site to ensure consistency among its multilingual portions. The result of this process is a Content Management System, which internally stores multilingual content and page structure.
- Before creating the Content Management System the Hyperlinks analyzer is applied. The result of this module is a list of ‘missing links’ able to make the hyperlinks structure of the target Web application consistent.
- A preliminary evaluation of the Hyperlinks analyzer has been conducted on a sample of 10 multilingual Web sites: in 8 cases out of 10 at least an internal inconsistency or an external one is present.
- As a future work we plan to conduct further experiments on dynamic Web applications.