iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora

Hernani Costa, Gloria Corpas Pastor and Miriam Seghiri
{hercos,gcorpas,seghiri}@uma.es

LEXYTRAD, University of Malaga
Malaga, Spain

Introduction

• The interest in mono-, bi- and multilingual corpora is vital in many research areas, such as:
  - terminology and specialised language
  - automatic and assisted translation
  - language teaching
  - natural language processing
  - amongst other research areas

• Particularly in translation, their benefits have been demonstrated by various authors [1, 2, 3, 4]

Comparable Corpora

• It is already a fact that using comparable corpora [5] is the solution for the lack of sufficient/up-to-date parallel corpora and linguistic resources, specially for narrow domains and poorly-resourced languages

• Some of the advantages of using comparable corpora are the following:
  - objectivity
  - reusability
  - multiplicity and applicability of uses
  - easy handling and quick access to large volume of data

Existing Corpora Compilation Solutions and their limitations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- compilation tools are scarce or proprietary</td>
<td>- built to compile one monolingual corpus at a time</td>
<td></td>
</tr>
<tr>
<td>- simplistic with limited features</td>
<td>- or do not cover the entire compilation process (i.e. they do not allow managing and exploring both parallel and multilingual comparable corpora)</td>
<td></td>
</tr>
<tr>
<td>- built to compile one monolingual corpus at a time</td>
<td>- addresses some limitations in current solutions, such as: the use of more than one boolean operator when creating search query strings</td>
<td></td>
</tr>
<tr>
<td>- or do not cover the entire compilation process (i.e. they do not allow managing and exploring both parallel and multilingual comparable corpora)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

iCompileCorpora

Manual

• Represents the option of compiling monolingual and multilingual corpora

• Allows for the manual upload of documents from a local or remote directory

Semi-automatic

• Permits the exploitation of both mono- and multilingual corpora mined from the Internet

• Addresses some limitations in current solutions, such as: the use of more than one boolean operator when creating search query strings

Semi-automatic CLIR

• Address the demand for multilingual corpora by taking advantage of CLIR techniques

• Allows for the retrieval of relevant information written in a language different to the one semi-automatically retrieved by the semi-automatic layer

iCompileCorpora Layered Model

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union’s Framework Programme (FP7/2007-2013) under REA grant agreement n° 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. n° FFI2012-38881, 2012-2015), and the R&D Project for Excellence TERMITUR (ref. n° HUM2754, 2014-2017).

References

http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.