Overview

• Automatic identification of reference fields in Digital Humanities (DH) data
• On-going project dealing with articles and blogs published on CLEO’s OpenEdition platform
• Automatic recognition of reference fields
• Automatic link creation between related articles and blogs

Bibliographical references in OpenEdition

http://www.openedition.org
http://www.revues.org

Model

Conditional Random Fields (Lafferty et al. 2001, ICML)

\[
P(y|x) = \frac{1}{Z(x)} \exp \left( \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t) \right),
\]

\( y = y_1 \ldots y_T \) : a label sequence
\( x = x_1 \ldots x_T \) : an input sequence
\( \theta = \{ \theta_k \} \in \mathbb{R}^K \) : parameter vector
\( f_k(y_t, y_{t-1}, x_t) \) : feature functions

Learning data example

<table>
<thead>
<tr>
<th>Reference data</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographical references in OpenEdition</td>
<td>Conditional Random Fields (Lafferty et al. 2001, ICML)</td>
</tr>
</tbody>
</table>

Experiments

Corpus in three different difficulty levels

• Manually annotated Revues.org data with TEI guidelines.
• For reuse of data, detailed annotation.

<table>
<thead>
<tr>
<th>level</th>
<th># references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus 1</td>
<td>715 references</td>
</tr>
<tr>
<td>Corpus 2 (footnote)</td>
<td>1147 (bibl.) 385 (non-bibl.)</td>
</tr>
<tr>
<td>Corpus 3 (context)</td>
<td>553 (context, note) 449 (only author, date)</td>
</tr>
</tbody>
</table>

• More 40 models (changing features and tokenization criteria)
• Both automatic and manual evaluation
• Insertion of external resources as features (author list etc.)
• Mallet software by McCallum : http://mallet.cs.umass.edu/

Results

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Accuracy</th>
<th>Precis.</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus 1</td>
<td>77.74</td>
<td>93.15</td>
<td>90.20</td>
</tr>
<tr>
<td>Corpus 2</td>
<td>75.68</td>
<td>52.17</td>
<td>17.91</td>
</tr>
<tr>
<td>Corpus 3</td>
<td>65.22</td>
<td>96.97</td>
<td>84.21</td>
</tr>
</tbody>
</table>

Future work

• Effective integration of external resources (post-processing etc.)
• CRF model construction on Corpus 2 and Corpus 3.
• Proper name disambiguation

Conceptual structure of Bilbo system

---

Young-Min Kim
LIA, University of Avignon
84911 Avignon, France
young-min.kim@univ-avignon.fr

Patrice Bellot
LSIS, Aix-Marseille University
13397 Marseille, France
patrice.bellot@lsis.org

Elodie Faath, Marin Dacos
CLEO, Centre for Open Electronic Publishing
13331 Marseille, France
{elodie.faath, marin.dacos}@revues.org

Young-Min Kim
LIA, University of Avignon
84911 Avignon, France
young-min.kim@univ-avignon.fr

Patrice Bellot
LSIS, Aix-Marseille University
13397 Marseille, France
patrice.bellot@lsis.org

Elodie Faath, Marin Dacos
CLEO, Centre for Open Electronic Publishing
13331 Marseille, France
{elodie.faath, marin.dacos}@revues.org

Patrice Bellot
LSIS, Aix-Marseille University
13397 Marseille, France
patrice.bellot@lsis.org

Elodie Faath, Marin Dacos
CLEO, Centre for Open Electronic Publishing
13331 Marseille, France
{elodie.faath, marin.dacos}@revues.org