CLRL: Feature Engineering for Cross-Language Record Linkage

Öyku Özlem Çakal
Technische Universität Berlin
o.cakal@campus.tu-berlin.de

Mohammad Mahdavi
Technische Universität Berlin
mahdavilahijani@tu-berlin.de

Ziawasch Abedjan
Technische Universität Berlin
abedjan@tu-berlin.de

Research Question
❑ How to link the records in a cross-language setting?
▪ Where each input dataset is in a different language

Challenges
❑ Simple translation of datasets may not work
▪ Ambiguity in translation
▪ Out-of-vocabulary (OOV) terms

Motivation Example
❑ Linking the movie “Forbidden Planet” to its German equivalent “Alarm im Weltall” is not trivial
▪ They have no lexical similarity
▪ The translated title “Alarm in the Universe” is still not similar to “Forbidden Planet”

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Jahr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Heat</td>
<td>1995</td>
</tr>
<tr>
<td>2</td>
<td>Forbidden Planet</td>
<td>1956</td>
</tr>
</tbody>
</table>

The Workflow of CLRL

OOV Term Expansion
❑ Morphological checking
  ▪ E.g., compound word “firstsight”
❑ Spell checking
  ▪ E.g., typo “researcch”
❑ Ostrich policy
  ▪ E.g., named entity “Lebowskii”

Feature Generation
❑ Monolingual similarity features
  ▪ Typical lexical measures
❑ Multilingual similarity features
  ▪ Via cross-language word embedding models [4]

Cross-Language Word Embedding Models...
❑ ... map words of different languages into a shared multidimensional space
❑ ... assign similar vectors to cross-lingual similar words, such as “dog” and “hund”

Mean Vector Similarity: Pick the similarity of mean vectors for two data cells
Max Vector Similarity: Pick the similarity of the most similar word pair for two data cells
Optimal Alignment Similarity: Pick the mean similarity of word pairs when the best 1-to-1 word alignment is computed for two data cells
Max Alignment Similarity: Pick the mean similarity of word pairs when each word is aligned to its most similar word in the other data cell

Experimental Setup
❑ 3 Baselines
  ▪ Magellan [1] (M)

❑ 6 Datasets
  ▪ Universities
  ▪ Movies
  ▪ Wikipedia Titles

<table>
<thead>
<tr>
<th>Name</th>
<th>Language</th>
<th>#Rows</th>
<th>#Common Attributes</th>
<th>#Actual Linked Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universities</td>
<td>English</td>
<td>8758</td>
<td>16</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>3957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Movies</td>
<td>English</td>
<td>1273</td>
<td>14</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>15334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikipedia Titles</td>
<td>English</td>
<td>1976</td>
<td>2</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>2130</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evaluation

Effectiveness of CLRL

References

Source Code
Our prototype is available online: https://github.com/bigdama/clrl

Acknowledgement
This work has been partially funded by the German Ministry for Education and Research as BTRC II (01IS18025A).

ID Name Year ID Name Jahr
1 Heat 1995 1 Alarm im Weltall 1956
2 Forbidden Planet 1956 2 Der Pate 1972