Using Global Constraints and Reranking to Improve Cognates Detection

Michael Bloodgood
Department of Computer Science
The Ohio State University

Benjamin Strauss
Computer Science and Engineering Dept.
The Ohio State University

Introduction
Background: Cognates detection is the task of identifying words across languages that have a common origin. Cognates are used for protolanguage reconstruction and cross-language dictionary lookup. Cognates can also improve the quality of machine translation, word alignment, and bilingual lexical induction.

Current Solution: Create a score matrix for all word pairs based on weighted combinations of component scores. The component scores are computed on the basis of word context information, word frequency information, temporal information, word burstiness information, and phonetic information.

Our Contribution: We propose a new algorithm for rescoring the matrix by taking into account scores assigned to other word pairs. Precision and recall are improved by incorporating global constraints and reranking.

Motivation

<table>
<thead>
<tr>
<th>Portuguese</th>
<th>...</th>
<th>cozinhar</th>
<th>...</th>
<th>andar</th>
</tr>
</thead>
<tbody>
<tr>
<td>caminar</td>
<td>.80</td>
<td>.40</td>
<td>.70</td>
<td>.70</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>cocinar</td>
<td>.99</td>
<td>.99</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Blue - Score from Initial Score Matrix

(1) Reverse Rank

Green - Score after Reverse Rank

Black - Score after Forward Rank

General Task

Let \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \). Extract \((x, y)\) pairs such that \((x, y)\) are in some relation \( R \).

General Algorithm

Score Matrix

\[
Score(x, y) = \begin{bmatrix} s_{x_1,y_1} & \cdots & s_{x_1,y_n} \\ \vdots & \ddots & \vdots \\ s_{x_n,y_1} & \cdots & s_{x_n,y_n} \end{bmatrix}
\]  

Rescoring

Reverse Rank

\[
reverse\_rank(x, y) = [x \in X | s_{x,y} \geq s_{x,z} \]  

Score RR

\[
score_{RR}(x, y) = s_{x,y} / reverse\_rank(x, y)
\]

Forward Rank

\[
forward\_rank(x, y) = [y \in Y | s_{x,y} \geq s_{z,y}]
\]

Combining Reverse Rank and Forward Rank

1-Step Approach

\[
score_{RR, FR, 1step}(x, y) = s_{x,y} / product
\]

2-Step Approach

\[
score_{RR, FR, 2step}(x, y) = reverse\_rank(x, y) \times forward\_rank(x, y)
\]

Maximum Assignment

We used the Hungarian Algorithm to optimize:

\[
\max_{Z, x, y \in Y} \sum_{(x,y) \in Z} score(x, y)
\]

s.t. 

\[
(x,y) \in Z \Rightarrow (x', y') \notin Z, \forall k \neq i
\]

\[
(x,y) \in Z \Rightarrow (x, y') \notin Z, \forall k \neq j
\]

Computation of Initial Score Matrix

Lemmatization

- English - NLTK WordNetLemmatizer [Bird et al., 2009]
- French, German, Spanish - TreeTagger [Schmid, 1994]

Word Context Information

- 2012 Google 5-gram corpus for English, French, German, and Spanish [Michel et al., 2010]

Frequency Information

- Computed using the same corpora as Word Context Information.

Temporal Information

- Computed using the following corpora:
  - English - English Gigaword Fifth Edition
  - French - French Gigaword Fifth Edition
  - Spanish - Spanish Gigaword Fifth Edition
  - German - Web crawling and extracting news articles from

Word Burstiness Information

- Computed using the same corpora as Temporal Information.

Phonetic Information

- Computed using a measurement based on Normalized Edit Distance (NED).

Combining Information Sources

For each candidate cognate pair \((x, y)\), its final score is:

\[
score(x, y) = \sum_{m, metric} w_m score_{m}(x, y)
\]

Experiments

We used the following language pairs:

- French-English
- German-English
- Spanish-English

We used the following testing conditions:

- Small Data (Traditional)
- Large Data (New)

We used the following as our baseline:

- Initial Score Matrix without any rescoring

Using Global Constraints to Rescore - Results

Figure: Precision-Recall Curves for French-English (small data)

Table: French-English Performance (small data)

<table>
<thead>
<tr>
<th>Method</th>
<th>Max F1</th>
<th>11-point IAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>54.92</td>
<td>50.99</td>
</tr>
<tr>
<td>RR</td>
<td>62.94</td>
<td>59.62</td>
</tr>
<tr>
<td>RR, FR, 1step</td>
<td>68.35</td>
<td>64.42</td>
</tr>
<tr>
<td>RR, FR, 2step</td>
<td>69.72</td>
<td>67.29</td>
</tr>
</tbody>
</table>

Figure: Precision-Recall Curves for French-English (large data)

Table: French-English Performance (large data)

<table>
<thead>
<tr>
<th>Method</th>
<th>Max F1</th>
<th>11-point IAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>55.08</td>
<td>51.35</td>
</tr>
<tr>
<td>RR</td>
<td>60.88</td>
<td>58.79</td>
</tr>
<tr>
<td>RR, FR, 1step</td>
<td>65.87</td>
<td>63.55</td>
</tr>
<tr>
<td>RR, FR, 2step</td>
<td>65.76</td>
<td>65.26</td>
</tr>
</tbody>
</table>

Conclusion

We presented new methods for rescoring a matrix of initial scores and new large data testing conditions for evaluation. Our new methods are complementary to existing state-of-the-art methods, easy to implement, computationally efficient, and effective in improving performance.

Acknowledgment

We would like to thank TCNJ students Ethan Kochis and Garrett Beatty for their help with building this poster.

Bibliography

- Bloodgood & Strauss (2010)