Introduction

The task of discovering stylistic elements of Charles Dickens has been addressed by Tabata [1], using Random Forest (RF) classification [2], by comparing Dickens to Wilkie Collins, as well as a larger set of contemporary authors from the 18th and 19th century. RF is an ensemble learning technique that averages feature importance over a large number of trees generated from the data. Here, we set ourselves the same task, but using a simple statistical measure to compute stylistic elements of “Dickens vs. Collins” and “Dickens vs. World”. In the absence of a gold standard, we evaluate heuristically, using separability in clustering.

Evaluation

We test whether the selected features can separate authors by clustering on the test set using representative and distinctive terms (i.e. Case I of Distinctiveness). We evaluate clustering using the Adjusted Rand Index [3], where 0 is the expected (chance) value and 1 perfect overlap with (a gold) standard.

To detect stylistic markers of authors, such as Dickens, we propose Representativeness (RD) [4], for revealing those features most consistent for Dickens while also being distinctive with respect to another author.

- Representative features: used consistently either frequently or infrequently by the author.
- Distinctive features: used differently compared to another author.

Table 1: Clustering using discriminatory terms on test sets based on 5-fold cross-validation.

<table>
<thead>
<tr>
<th>Test set</th>
<th>Rand Index</th>
<th>Test set</th>
<th>Rand Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.49</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>0.22</td>
<td>5</td>
<td>0.20</td>
</tr>
</tbody>
</table>

References