Language identification of names with SVMs

ADITYA BHARGAVA AND GRZEGORZ KONDRAK
UNIVERSITY OF ALBERTA

NAACL-HLT 2010
JUNE 3, 2010
Outline

- Introduction: task definition & motivation
- Previous work: character language models
- Using SVMs
- Intrinsic evaluation
  - SVMs outperform language models
- Applying language identification to machine transliteration
  - Training separate models
- Conclusion & future work
Task definition

- Given a name, what is its language?
- Same script (no diacritics)

Beckham  ➔  English
Brillault ➔  French
Velazquez ➔  Spanish
Friesenbichler ➔  German
Motivation

- Improving letter-to-phoneme performance (Font Llitjós and Black, 2001)
- Improving machine transliteration performance (Huang, 2005)
- Adjusting for different semantic transliteration rules between languages (Li et al., 2007)
Previous approaches

- Character language models (Cavnar and Trenkle, 1994)
  - Construct models for each language, then choose the language with the most similar model to the test data
  - 99.5% accuracy given >300 characters & 14 languages
- Given 50 bytes (and 17 languages), language models give only 90.2% (Kruengkrai et al., 2005)
- Between 13 languages, average F1 on last names is 50%; full names gives 60% (Konstantopoulos, 2007)
- Easier with more dissimilar languages: English vs. Chinese vs. Japanese (same script) gives 94.8% (Li et al., 2007)
Using SVMs

• **Features**
  - Substrings (n-grams) of length n for n=1 to 5
    - Include special characters at the beginning and the end to account for prefixes and suffixes
  - Length of string

• **Kernels**
  - Linear, sigmoid, RBF
  - Other kernels (polynomial, string kernels) did not work well
Evaluation: Transfermarkt corpus

- **European national soccer player names** (Konstantopoulos, 2007) from 13 national languages
  - ~15k full names (average length 14.8 characters)
  - ~12k last names (average length 7.8 characters)

- **Noisy data**
  - e.g. Dario Dakovic born in Bosnia but plays for Austria, so annotated as German
Evaluation: Transfermarkt corpus

Accuracy

Language models
Linear SVM
RBF SVM
Sigmoid SVM

Last names
Full names
## Evaluation: Transfermarkt corpus

<table>
<thead>
<tr>
<th></th>
<th>cs</th>
<th>da</th>
<th>de</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>it</th>
<th>nl</th>
<th>no</th>
<th>pl</th>
<th>pt</th>
<th>se</th>
<th>yu</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>19</td>
<td>0</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>0.33</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>27</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.46</td>
</tr>
<tr>
<td>de</td>
<td>4</td>
<td>2</td>
<td>183</td>
<td>12</td>
<td>2</td>
<td>11</td>
<td>2</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>0.72</td>
</tr>
<tr>
<td>en</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>69</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.62</td>
</tr>
<tr>
<td>es</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>4</td>
<td>25</td>
<td>7</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0.31</td>
</tr>
<tr>
<td>fr</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>10</td>
<td>5</td>
<td>41</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0.43</td>
</tr>
<tr>
<td>it</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td>nl</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>36</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.42</td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0.36</td>
</tr>
<tr>
<td>pl</td>
<td>2</td>
<td>0</td>
<td>13</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0.68</td>
</tr>
<tr>
<td>pt</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>se</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>23</td>
<td>4</td>
<td>0.43</td>
</tr>
<tr>
<td>yu</td>
<td>3</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>84</td>
<td>0.76</td>
</tr>
</tbody>
</table>

| Precision | 0.53 | 0.68 | 0.55 | 0.58 | 0.40 | 0.39 | 0.59 | 0.59 | 0.46 | 0.70 | 0.27 | 0.74 | 0.74 |
Evaluation: CEJ corpus

- Chinese, English, and Japanese names (Li et al., 2007)
  - ~97k total names, average length 7.6 characters
- Demonstrates a higher baseline with dissimilar languages
- Linear SVM only (RBF and sigmoid were slow)
Language origin knowledge may help machine transliteration systems pick appropriate rules

To test, we manually annotated data

- English-Hindi transliteration data set from the NEWS 2009 shared task (Li et al., 2009; MSRI, 2009)
- 454 “Indian” names, 546 “non-Indian” names
- Average length 7 characters

SVM gives 84% language identification accuracy
Application to machine transliteration

- Basic idea: use language identification to split data into two language-specific sets
- Train two separate transliteration models (with less data per model), then combine
- We use DirecTL (Jiampojamarn et al., 2009)
- Baseline comparison: random split
- Three tests:
  - DirecTL (Standard)
  - DirecTL with random split (Random)
  - DirecTL with language identification–informed split (LangID)
Application to machine transliteration
Conclusion

• Language identification of names is difficult
  ○ SVMs with n-grams as features work better than language models

• No significant effect on machine transliteration
  ○ But there does seem to be some useful information
Future work

- Web data
- Other ways of incorporating language information for machine transliteration
  - Direct use as a feature
  - Overlapping (non-disjoint) splits
Questions?