1. Overview

Language identification is an important pre-processing step for many NLP systems. We developed software for the language identification of electronic text documents.

- Language Corpus and Input Document based on Wikipedia
- Data of 76 Languages are used
- The system structure permits expansion of training corpus with further data
- Input and output occur with Web-Interface
- Approaches: Step-By-Step, Ranking, and Vector Space Model
- Classification Methods: probability distribution (Bayes Decision Rule), Ad-hoc Ranking (Out-Of-Place Measure) and Vector Space Model

3. Approaches

Step-By-Step
- Determination of the most probable language with Bayes Decision Rule
- Features: Tokens, Bi-grams

Ad-hoc Ranking
- Comparison of the document model with the language model by Out-Of Place Measure
- Features: N-grams (2-4)

Vector Space Model
- Determination of similarity between the document model vector and the language model vectors by Cosine Distance
- Features: N-grams (2-4) + Tokens

4. Web Interface

5. Evaluation

Average performance using just English, German, Danish and Dutch

<table>
<thead>
<tr>
<th>Chunk size</th>
<th>Step by Step</th>
<th>Vector Model</th>
<th>Ranking</th>
<th>Simple Ranking</th>
<th>Simple Vector Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>95.2</td>
<td>94.2</td>
<td>92.2</td>
<td>96.2</td>
<td>98.7</td>
</tr>
<tr>
<td>25</td>
<td>99.7</td>
<td>99.2</td>
<td>99.7</td>
<td>96.2</td>
<td>99.7</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>99.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>99.7</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>99.7</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

6. Conclusion

- 200 KB test data and 500/1000 KB training data for each language are used
- 1000 KB data produces better results as 500 KB
- Method based on combination of features did better than methods that employed single features
- Step-By-Step method performed better for all chunk sizes
- Success of the identification is indirectly proportional to the number of languages in corpus; the more the languages the worse the results
- Documents often include words in more than one language, which complicates the correct language identification

7. References

- Language Identification With Confidence Limits (David Elworthy, 1998)
- N-Gram-Based Text Categorization (William B.Cavnar and John M. Treacle, 1994)
- Linguini: Language Identification for Multilingual Documents (Prager, J. M., 1999)