A Hybrid Approach for Transliterated Word-Level Language Identification: CRF with Post Processing Heuristics

Authors: Somnath Banerjee+, Aniruddha Roy+, Alapan Kuila+, Sudip Kumar Naskar+, Sivaji Bandyopadhyay+, and Paolo Rosso#

+Jadavpur University, India & #Technical University of Valencia, Spain

Team: JU-NLP-LAB

Presented By-

Somnath Banerjee
Overview

- Introduction.
- Definition of task.
- Resources and Dataset
- Hybrid System description
- Result
- Conclusion
- References
Introduction

- Participated in Subtask-1 of *Transliterated Search* for Bengali Language.
- Employed a *Hybrid approach* for word-level language (WLL) identification.
- Two transliteration systems were built to transliterate detected Bangla words.
Resources and Dataset

Dataset:

- The training and the test data have been constructed by using manual and automated techniques.
- The training dataset consists of 800 lines.
- The test set contains 1000 sentences.

Resources:

- English word frequency list.
- Bangla word frequency list
- Bangla word transliteration pairs dataset.
System Description

- Language Identification Task
  - Machine Learning Approach
  - Post Processing with Heuristics
- Transliteration Task
Language Identification System
Machine Learning Approach

Feature Used:

- Character N-Gram.
- Symbol character.
- Links
- Presence of digits.
- Word suffix.
- Contextual Probability.
Character n-grams

- Few studies [9, 5] successfully used the character n-gram feature and obtained reasonable results.
- We used this feature from character unigrams to 5-grams.
Symbol Characters

- A word may start with a symbol, such as #Engineering, @India, etc.
- Also a symbol may appear within the word itself, such as a****a, kanka-r etc.
- Sometimes the entire word is built up of a symbol, e.g. “, ?.

\[
\text{has-symbol(word)} = \begin{cases} 
1, & \text{if word contain any symbol.} \\
0, & \text{Otherwise}
\end{cases}
\]
Links

- used as a binary feature.

\[
\text{is-link(word)} = \begin{cases} 
1 & \text{, if word contain link} \\
0 & \text{, Otherwise}
\end{cases}
\]
Presence of Digit

- digit(s) may mean different things in the chat dialogue.
  
  E.g., 'gr8', 'b4'.

- Used as a binary feature.

\[
\text{has-digit(word)} = \begin{cases} 
1, & \text{If it has digit} \\
0, & \text{Otherwise}
\end{cases}
\]
Word Suffix

- Language dependent features
- Prepared a small suffix-list (10 entries) under human supervision from studying an archive (10 documents) of an online Bangla newspaper.
- Used the feature as a binary set.

\[
\text{has-suffix}(\text{word}) = \begin{cases} 
1 & \text{if the word contains any suffix} \\
0 & \text{Otherwise}
\end{cases}
\]
Contextual Feature

- Words like `take`, `are`, `pore`, `bad` are truly ambiguous words with respect to the WLL identification task as they are valid English words as well as backward transliterations of valid Bangla words.
  - E.g.,
    - Mama take this badge off of me.
    - Ami take (Him) boli je ami bansdronir kichu agei thaki.

- Probability of the previous and the next word belonging to a particular language {E/B/H/O}.
Contextual Probability

- The probability is calculated as
  
  \[ P_{\text{tag}}(w) = \frac{F_{\text{tag}}(w)}{F(w)}, \]
  
  where tag \( \in \{E,B,H,O\} \)

- We used a smoothed version
  
  \[ P_{\text{tag}}(w) = \frac{F_{\text{tag}}(w) + 1}{F(w) + N} \]

Where,

- \( F_{\text{tag}}(w) = \) frequency of the word \( W \) belonging to tag
- \( F(w) = \) Frequency of the word
- \( N = \) total number of words in the training corpus.
WLL Classifier

- Used Conditional Random Field (CRF) to build the model for WLL identification classifier.
- Used CRF++ toolkit\(^1\) implementation of CRF.

\(^1\)http://crfpp.googlecode.com/svn/trunk/doc/index.html
Post Processing
## Post Processing

<table>
<thead>
<tr>
<th>Rule</th>
<th>End with</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-1</td>
<td>ed</td>
<td>decided, reached</td>
</tr>
<tr>
<td>Rule-2</td>
<td>ly</td>
<td>thoughtfully, anxiously, unfriendly</td>
</tr>
<tr>
<td>Rule-3</td>
<td>ing</td>
<td>evening, kissing, playing</td>
</tr>
<tr>
<td>Rule-4</td>
<td>‘s</td>
<td>women's, uncle's</td>
</tr>
<tr>
<td>Rule-5</td>
<td>‘t</td>
<td>don’t, isn’t, wouldn’t</td>
</tr>
<tr>
<td>Rule-6</td>
<td>‘ll</td>
<td>I’ll, It’ll, he’ll, you’ll</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule</th>
<th>Start with</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-7</td>
<td>O’</td>
<td>O’clock, O’Keefe</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mis-spelled</th>
<th>Applied Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>lvly, xactly, physicaly</td>
<td>Rule-2</td>
</tr>
<tr>
<td>lukiing, nthing, njoying</td>
<td>Rule-3</td>
</tr>
<tr>
<td>rn’t, cudn’t</td>
<td>Rule-5</td>
</tr>
</tbody>
</table>
### Post Processing

<table>
<thead>
<tr>
<th>Rule</th>
<th>Condition</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-8</td>
<td>Special symbol (e.g. #,$,&amp; )</td>
<td>Others</td>
</tr>
<tr>
<td>Rule-9</td>
<td>English dictionary + Contextual probability &gt; threshold value (0.08)</td>
<td>English</td>
</tr>
<tr>
<td>Rule-11</td>
<td>Links</td>
<td>Others</td>
</tr>
</tbody>
</table>

**Rule-10: Observations**

Repetition of a character more than twice

<table>
<thead>
<tr>
<th>Position</th>
<th>Example</th>
<th>Most Likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>end</td>
<td>torengeee, plzzzzz</td>
<td>English/Hindi</td>
</tr>
<tr>
<td>middle</td>
<td>kisssob, oneeek</td>
<td>Bengali</td>
</tr>
<tr>
<td>middle + end</td>
<td>muuuuaaahhhhhhhh</td>
<td>English</td>
</tr>
</tbody>
</table>
Transliteration System
Approach-1: 2 PASS MODEL

Word to be Transliterated

Character Chunk Model
If Length=Even || Length=5, divide the word into blocks of 2 chars
Otherwise, divide the words into block of 3 chars

Model-1

Character Model
All words are divided into single chars

Model-2

Decoded O/P

Decoded O/P

- After two pass, all the words are transliterated.
Approach-2: Syllable-level transliteration

• Break words into transliteration units (TU) following the heuristic used in (Ekbal et al., 2006).
  – English words broken into TUs that have the pattern C*V*.
  – Bengali words broken into TUs that have the pattern C+M.
## Results

<table>
<thead>
<tr>
<th>Token level language accuracy</th>
<th>Language</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangla</td>
<td>0.866</td>
<td>0.935</td>
<td>0.899</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>0.944</td>
<td>0.899</td>
<td>0.920</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Token level Transliteration</th>
<th>Run</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-1</td>
<td>0.033</td>
<td>0.572</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>Run-2</td>
<td>0.019</td>
<td>0.338</td>
<td>0.037</td>
<td></td>
</tr>
</tbody>
</table>

| Language Identification Accuracy | 0.905 |
## Results

<table>
<thead>
<tr>
<th>Other Performance Metrics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQMF All (No Translit.)</td>
<td>0.444</td>
</tr>
<tr>
<td>EQMF without NE (No Translit.)</td>
<td>0.548</td>
</tr>
<tr>
<td>EQMF without MIX (No Translit.)</td>
<td>0.444</td>
</tr>
<tr>
<td>EQMF without NE &amp; MIX (No Translit.)</td>
<td>0.548</td>
</tr>
<tr>
<td>EQMF All Run-1</td>
<td>0.005</td>
</tr>
<tr>
<td>EQMF All Run-2</td>
<td>0.004</td>
</tr>
<tr>
<td>EQMF without NE: Run-1</td>
<td>0.007</td>
</tr>
<tr>
<td>EQMF without NE: Run-2</td>
<td>0.004</td>
</tr>
<tr>
<td>EQMF without MIX: Run-1</td>
<td>0.005</td>
</tr>
<tr>
<td>EQMF without MIX: Run-2</td>
<td>0.004</td>
</tr>
<tr>
<td>EQMF without NE &amp; MIX: Run-1</td>
<td>0.007</td>
</tr>
<tr>
<td>EQMF without NE &amp; MIX: Run-2</td>
<td>0.004</td>
</tr>
<tr>
<td>ETPM: Run-1</td>
<td>227/364</td>
</tr>
<tr>
<td>ETPM: Run-2</td>
<td>134/364</td>
</tr>
</tbody>
</table>
Conclusion & Future Work

- **Conclusion**
  - We presented a hybrid approach to address the automatic WLL identification problem.
  - Simple post-processing heuristics enhances the overall performance of the WLL system.
  - Two variants of the transliteration systems were developed.

- **Future work:**
  - Introduce more features for the machine learning model.
  - New post-processing heuristics for the WLL identification task.
  - Make use of contextual information for transliteration.
References


References


Thanks