Ensemble Classifier based Approach for Code-Mixed Cross-Script Question Classification

Team : IINTU

Debjyoti Bhattacharjee
School of Computer Science and Engineering
Nanyang Technological University
Singapore

Paheli Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India
OUTLINE OF THE PRESENTATION

• Mixed Script Information Retrieval (MSIR)
• Question Classification in Code-Mixed data
• Proposed Approach
• Experimental Setup
• Results
• Conclusion and Future Work
Mixed-Script/Code-Mixed Data

Adjust Kijiye

Ek Chance Milega?

Breadyatra!

Bahut Tension Hai

"Bahut Garmi Hai, Yaar!"

English hamari sanskriti ka hissa hai!

Amul
Tasty in any language!
Both documents and queries are in more than one scripts

**Transliterated** from native script (Devnagari for Hindi) to foreign script (Roman)

Define MSIR formally¹:

- Natural languages $L = \{l_1, l_2, \ldots, l_n\}$
- Scripts $S = \{s_1, s_2, \ldots, s_n\}$ such that $s_i$ is the native script for language $l_i$
- Word $w_i = < l_i, s_j >$
- $i = j$, native script, else transliterated

---

¹Gupta et. al., Query Expansion for Mixed-Script Information Retrieval, SIGIR 2014
Why MSIR?

- Users now opt to write in their native language rather than English
- Shortcoming: Font-encoding issues, English keyboard
- Write in the Roman Script by transliteration
• Question Answering
  – Find concise and accurate answer to a given question

• Question Classification
  – Subtask of Question Answering
  – Determine the type of answer for a question

• Categorize a question in to a set of classes and deal with each class for answering
• Mixing of the languages English and Bengali
• Set of questions $Q = \{q_1, q_2, \ldots, q_n\}$
• Each question $q = \langle w_1 \ w_2 \ \ldots \ \ w_n \rangle$
  $\quad - \ w_i = $ English word or transliterated Bengali
• Set of classes $C = \{c_1, c_2, \ldots, c_m\}$
• Classify question $q_i$ to a class $c_j$
Question Classification in Mixed-Script

Kharagpur theke Howrah car fare koto?

- Bengali
- English

- Distance
- Temporal
- Money
- Location
Proposed Approach

• Each question is represented as a 2000 dimensional binary vector
  – $i^{th}$ component $\leftrightarrow$ the $i^{th}$ most frequent word
• Train classifiers
  – Random Forests (RF)
  – One-Vs-Rest (OvR)
  – k-Nearest Neighbours (kNN)
• Ensemble of the classifiers
  – Majority Vote
  – Else, a random label
• Retraining
  – From the test set, pick up 90% of the samples (by replacement) which had the same label for all the 4 classifiers
  – New training = Original Training Set + Sampled Test Set
Random Forest (RF)

- Ensemble learning method
- Fits a number of decision trees on various sub-samples of the dataset
- Use averaging to improve the predictive accuracy and control over-fitting
One-vs-all (one-vs-rest):

- Fits one classifier per class $i$ to predict $p(\text{class}=i \mid x, \theta)$
- Test sample, pick the class $i$ that has the maximum probability
- Each classifier is trained with the entire dataset
- Most commonly used strategy for multiclass classification
**K-Nearest Neighbours (kNN)**

- Majority class vote of its neighbours
- Being a non-parametric method, it is often successful in classification situations where the decision boundary is very irregular
- Simple classifier
Ensemble Classifier

Question Vector
[1 0 1 0 0 ...... 1 0 1]

RF
Class: TEMP

OvR
Class: NUM

k-NN
Class: TEMP

Majority Vote

Final Class: TEMP
**ENSEMBLE CLASSIFIER**

Question Vector

\[ 1 \, 0 \, 1 \, 0 \, 0 \, \ldots \, 1 \, 0 \, 1 \]  

- **RF**  
  Class: TEMP

- **OvR**  
  Class: NUM

- **k-NN**  
  Class: MISC

---

Random

Final Class: NUM
Retraining

Original Training Data

Sample of Test data

New Training Data

New Classifier

Test Data
<table>
<thead>
<tr>
<th>CLASS</th>
<th>NO. OF QUESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person (PER)</td>
<td>55</td>
</tr>
<tr>
<td>Location (LOC)</td>
<td>26</td>
</tr>
<tr>
<td>Organization (ORG)</td>
<td>67</td>
</tr>
<tr>
<td>Temporal (TEMP)</td>
<td>61</td>
</tr>
<tr>
<td>Numerical (NUM)</td>
<td>45</td>
</tr>
<tr>
<td>Distance (DIST)</td>
<td>24</td>
</tr>
<tr>
<td>Money (MNY)</td>
<td>26</td>
</tr>
<tr>
<td>Object (OBJ)</td>
<td>21</td>
</tr>
<tr>
<td>Miscellaneous (MISC)</td>
<td>5</td>
</tr>
</tbody>
</table>
• *scikit-learn* toolkit of Python 3
• Training-Validation Split = 9:1
• No. of trees in RF = 100
• Classifier for OvR = Linear SVC
• No. of neighbours in kNN = 30
RESULTS

OVERALL PERFORMANCE

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>78.19</td>
</tr>
<tr>
<td>OvR</td>
<td>81.11</td>
</tr>
<tr>
<td>RF</td>
<td>83.33</td>
</tr>
<tr>
<td>EC</td>
<td>81.67</td>
</tr>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>PER</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>23</td>
</tr>
<tr>
<td>LOC</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>26</td>
</tr>
<tr>
<td>ORG</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td>NUM</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>29</td>
</tr>
<tr>
<td>TEMP</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td>MONEY</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td>DIST</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>22</td>
</tr>
<tr>
<td>OBJ</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>MSC</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
• Machine learning algorithms for code-mixed Bengali-English data
• Scalable to other code-mixed questions since it is not language dependent
• Incorporate feature engineering – syntactic and semantic features
• Apply other ML algorithms
• Experiment with multi-script data
This work is supported by the project


funded by the Ministry of Human Resource Development (MHRD), India
THANK YOU

Any Questions!