Query Word Labeling and Transliteration for Indian Languages:
IITP_T5 Shared Task system description

Shubham Kumar Deepak Kumar Gupta Dr. Asif Ekbal

Department of Computer Science & Engineering
Indian Institute of Technology Patna

1st December 2014
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Outline

1. Language Identification & Transliteration
   - Subtask 1
     - Query Word Labelling
     - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Suppose that $q: w_1 \ w_2 \ w_3 \ldots \ w_n$, is a query is written Roman script.

The words, $w_1 \ w_2$ etc., could be standard English words or transliterated from another language $L$.

Task is to label the words as E or L depending on whether it an English word, or a transliterated $L$-language word.

Perform back transliteration for each transliterated word.
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
In social media communication, multilingual speakers often switch between languages.

Now a days many Indian languages especially in social media is written using romanized script

<table>
<thead>
<tr>
<th>Input Query</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>sachin tendulkar ka last test match</td>
<td>[sachin]P [tendulkar]P ka\H last\E test\E match\E</td>
</tr>
<tr>
<td>Jagjeet Singh ki famous gazal</td>
<td>[Jagjeet]P [Singh]P ki\H famous\E gazal\H</td>
</tr>
<tr>
<td>mars orbiter mission isro</td>
<td>mars\E orbiter\E mission\E [isro]O</td>
</tr>
<tr>
<td>IIT Patna Mathematics Department</td>
<td>[IIT]O [Patna]O Mathematics\E Department\E</td>
</tr>
<tr>
<td>Malgudi days ka pahla episode</td>
<td>Malgudi\H days\E ka\H pahla\H episode\E</td>
</tr>
</tbody>
</table>

**Table 1**: Query Word Labelling
Outline

1 Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2 Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3 Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4 Conclusions
Transliteration

- It is the process of converting a word written in one language into another language, by preserving the sounds of the syllables in words.
- It used when original script is not available to write down a word in that script.
- Majority of the population still use their mother-tongue as the medium of communication.
- Back-transliteration is the backward process that finds the origin word from the transliterated word.
## Transliteration

<table>
<thead>
<tr>
<th>Input Query</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>sachin tendulkar ka last test match</td>
<td>[sachin]P [tendulkar]P ka\H=का last\E test\E match\E</td>
</tr>
<tr>
<td>Jagjeet Singh famous gazal</td>
<td>[Jagjeet]P [Singh]P famous\E gazal\H=गजल</td>
</tr>
<tr>
<td>mars orbiter mission isro</td>
<td>mars\E orbiter\E mission\E [isro]O</td>
</tr>
<tr>
<td>IIT Patna Mathematics Department</td>
<td>[IIT]O [Patna]O Mathematics\E Department\E</td>
</tr>
<tr>
<td>bharat ka australia daura</td>
<td>[bharat]L ka\H=का [australia]L daura\H=दौरा</td>
</tr>
</tbody>
</table>

**Table 2: Transliteration Labelling**
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Query Word Labelling

Language Identification: Develop the systems based on four different classifier namely **Support vector machine**, **Decision tree**, **Random forest** and **Random tree** and finally combine their outputs using a majority voting technique.

The different features which we used for classification are as follows:

1. Character n-gram
2. Gazetteer based feature
3. Context word
4. Word normalization
5. InitCap
6. InitPunDigit
7. DigitAlpha
8. Contains# symbol
Features

1. **Character n-gram**: extracted character n-grams of length one (unigram), two (bigram) and three (trigram).
2. **Context word**: used the contexts of previous two and next two words as features.
3. **Word normalization**: each capitalized letter is replaced by A, small by a and number by 0.
4. **Gazetteer based feature**: checked from the compiled list of Hindi, Bengali and English words from the training datasets.
5. **InitCap**: checks whether the current token starts with a capital letter.
6. **InitPunDigit**: defined a binary-valued feature that checks whether the current token starts with a punctuation or digit.
7. **DigitAlpha**: defined this feature in such a way that checks whether any token in the surrounding context is alphanumeric.
8. **Contains# symbol**: defined the feature that checks whether the word in the surrounding context contains the symbol #.
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Methodology

- Query Word Labelling
  - Language Identification
  - Named Entity Recognition & Classification (NERC)
  The task was to identify named entities (NEs) and classify them into the following categories: Person, Organization, Location and Abbreviation
  Develop the systems based on four different classifier namely Support vector machine, Decision tree, Random forest and Random tree.
  The different features which we used for NERC are as follows:
  1. Local context
  2. Character n-gram
  3. Prefix and Suffix
  4. Word normalization
  5. WordClassFeature
  6. Typographic features
Features

1. **Local context**: used the previous two and next two tokens as the features.

2. **Character n-gram**: used n-grams of length upto 5 as the features.

3. **Prefix and Suffix**: of fixed length 3 character sequences are stripped from each token.

4. **Word normalization**: same as we did for language identification.

5. **WordClassFeature**: normalized all the words following the process as mentioned above. Thereafter, consecutive same characters are squeezed into a single character.

6. **Typographic features**: implemented four features: AllCaps (whether the current word is made up of all capitalized letters), AllSmall (word is constructed with only uncapitalized characters), InitCap (word starts with a capital letter) and DigitAlpha (word contains digits and alphabets).
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Methodology

- Query Word Labelling
- Transliteration
  Two levels of decoding process
  1. Segmenting source and target language strings into TU.
     - Regular expression of source TU: C*V*
     - C: Consonants & V: Vowels
     - Regular expression of target TU: C+M
     - C: Consonants, vowels or conjuncts & M: vowel modifier or Matra
     - For example:
       pa | ha | la ↔ प | ह | ला
Defining appropriate mapping between the source and target TU. 

Three Models developed

Final probability is calculated by multiplying maximum probability that exists between a source TU and a target TU in each mapping

- **Model-I**
  - No context is considered
  
  \[ P(X,T) = \prod_{k=1}^{k} P(\langle x, t \rangle_k) \]

- **Model-II**
  - Next source TU considered as context
  
  \[ P(X,T) = \prod_{k=1}^{k} P(\langle x, t \rangle_k | x_{k+1}) \]

- **Model-III**
  - Considered next source TU and back-transliteration of previous source TU as context
  
  \[ P(X,T) = \prod_{k=1}^{k} P(\langle x, t \rangle_k | \langle x, t \rangle_{k-1}, x_{k+1}) \]

**P(X,T):** probability that T is the back-transliteration of X

- x: Source TU
- t: Target TU
- k: no. of mappings between source and target TU’s
Model I

<table>
<thead>
<tr>
<th>Source TU</th>
<th>Destination TU</th>
<th>Frequency</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>pa</td>
<td>प</td>
<td>80</td>
<td>0.80</td>
</tr>
<tr>
<td>pa</td>
<td>पा</td>
<td>15</td>
<td>0.15</td>
</tr>
<tr>
<td>pa</td>
<td>पै</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>ha</td>
<td>ह</td>
<td>70</td>
<td>0.70</td>
</tr>
<tr>
<td>ha</td>
<td>हा</td>
<td>22</td>
<td>0.22</td>
</tr>
<tr>
<td>ha</td>
<td>है</td>
<td>8</td>
<td>0.08</td>
</tr>
<tr>
<td>la</td>
<td>ला</td>
<td>65</td>
<td>0.65</td>
</tr>
<tr>
<td>la</td>
<td>ल</td>
<td>15</td>
<td>0.15</td>
</tr>
<tr>
<td>la</td>
<td>लै</td>
<td>20</td>
<td>0.20</td>
</tr>
</tbody>
</table>
pa | ha | la

\[
\begin{align*}
\text{pa} & \rightarrow \text{पा (0.80)} \\
\text{ha} & \rightarrow \text{हा (0.70)}
\end{align*}
\]

\[
\begin{align*}
\text{pa} & \rightarrow \text{पै (0.05)} \\
\text{ha} & \rightarrow \text{है (0.22)}
\end{align*}
\]

\[
\begin{align*}
\text{la} & \rightarrow \text{ला (0.65)} \\
\text{la} & \rightarrow \text{ल (0.15)} \\
\text{la} & \rightarrow \text{ले (0.20)}
\end{align*}
\]

\[
\begin{align*}
\text{pa} &= \max (0.80, 0.05, 0.15) = 0.80 \ (प)
\text{ha} &= \max (0.70, 0.22, 0.08) = 0.70 \ (ह)
\text{la} &= \max (0.65, 0.15, 0.20) = 0.65 \ (ला)
\end{align*}
\]

पहला
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Data-Sets for Training

- Query Word Labelling
  - Used data-sets of FIRE 2014.
  - consists of 20,658 and 27,969 tokens respectively for Hindi-English & Bangla-English mixed scripts

- Transliteration
  - Used data-sets of FIRE 2013.
  - collated 54791 transliterated Hindi words
  - collated 19582 transliterated Bangla words
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
Run Description

- **Run 1**
  - For language identification and NERC, constructed the ensembles using majority voting
  - Back-Transliterated tokens that were labeled as Hindi or Bangla

- **Run 2**
  - Constructed language identification by majority ensemble, and NERC by SMO
  - Back-Transliterated tokens that were labeled as Hindi or Bangla

- **Run 3**
  - Constructed language identification and NERC by SMO
  - Back-Transliterated tokens that were labeled as Hindi or Bangla
## Test set Results- Query Word Labelling

<table>
<thead>
<tr>
<th>Run ID</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
<th>EP</th>
<th>ER</th>
<th>EF</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-1</td>
<td>0.920</td>
<td>0.843</td>
<td>0.880</td>
<td>0.883</td>
<td>0.932</td>
<td>0.907</td>
<td>0.886</td>
</tr>
<tr>
<td>Run-2</td>
<td>0.922</td>
<td>0.843</td>
<td>0.881</td>
<td>0.884</td>
<td>0.931</td>
<td>0.907</td>
<td>0.886</td>
</tr>
<tr>
<td>Run-3</td>
<td>0.882</td>
<td>0.841</td>
<td>0.861</td>
<td>0.88</td>
<td>0.896</td>
<td>0.888</td>
<td>0.870</td>
</tr>
</tbody>
</table>

**Table-3** Result for language identification of Bangla-English

<table>
<thead>
<tr>
<th>Run ID</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
<th>EP</th>
<th>ER</th>
<th>EF</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-1</td>
<td>0.921</td>
<td>0.895</td>
<td>0.908</td>
<td>0.89</td>
<td>0.908</td>
<td>0.899</td>
<td>0.879</td>
</tr>
<tr>
<td>Run-2</td>
<td>0.921</td>
<td>0.893</td>
<td>0.907</td>
<td>0.89</td>
<td>0.908</td>
<td>0.899</td>
<td>0.878</td>
</tr>
<tr>
<td>Run-3</td>
<td>0.905</td>
<td>0.865</td>
<td>0.885</td>
<td>0.86</td>
<td>0.886</td>
<td>0.873</td>
<td>0.857</td>
</tr>
</tbody>
</table>

**Table -4** Results for language identification of Hindi-English
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
## Test set Results-Transliteration

**Table-5** Results of Transliteration for Bangla-English

<table>
<thead>
<tr>
<th>Run ID</th>
<th>EQMF ALL</th>
<th>TP</th>
<th>TR</th>
<th>TF</th>
<th>ETPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-1</td>
<td>0.005</td>
<td>0.039</td>
<td>0.574</td>
<td>0.073</td>
<td>228/337</td>
</tr>
<tr>
<td>Run-2</td>
<td>0.005</td>
<td>0.039</td>
<td>0.574</td>
<td>0.073</td>
<td>228/337</td>
</tr>
<tr>
<td>Run-3</td>
<td>0.005</td>
<td>0.038</td>
<td>0.582</td>
<td>0.071</td>
<td>231/344</td>
</tr>
</tbody>
</table>

**Table-6** Results of transliteration for Hindi-English

<table>
<thead>
<tr>
<th>Run ID</th>
<th>EQMF ALL</th>
<th>TP</th>
<th>TR</th>
<th>TF</th>
<th>ETPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-1</td>
<td>0.005</td>
<td>0.146</td>
<td>0.76</td>
<td>0.244</td>
<td>1933/2306</td>
</tr>
<tr>
<td>Run-2</td>
<td>0.004</td>
<td>0.146</td>
<td>0.76</td>
<td>0.244</td>
<td>1931/2301</td>
</tr>
<tr>
<td>Run-3</td>
<td>0.004</td>
<td>0.143</td>
<td>0.736</td>
<td>0.24</td>
<td>1871/2226</td>
</tr>
</tbody>
</table>
Outline

1. Language Identification & Transliteration
   - Subtask 1
   - Query Word Labelling
   - Transliteration

2. Methodology
   - Language Identification
   - Named Entity Recognition & Classification (NERC)
   - Transliteration

3. Results & Analysis
   - Data-sets
   - Results: Query Word Labelling
   - Results: Transliteration
   - Error Analysis

4. Conclusions
## Error Analysis

### Table-7 Language labeling errors. Here, H-Hindi, E-English, O-others

<table>
<thead>
<tr>
<th>Type</th>
<th>Words</th>
<th>Predicted</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short words</td>
<td>thrgh</td>
<td>H</td>
<td>E</td>
</tr>
<tr>
<td>Ambiguous words</td>
<td>the;ate</td>
<td>E;E</td>
<td>H;H</td>
</tr>
<tr>
<td>Erroneous words</td>
<td>implemnt</td>
<td>H</td>
<td>E</td>
</tr>
<tr>
<td>Mixed Numerals Words</td>
<td>2mar</td>
<td>O</td>
<td>B</td>
</tr>
</tbody>
</table>

### Table-8 Transliteration errors

<table>
<thead>
<tr>
<th>Type</th>
<th>Words</th>
<th>Predicted</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling Variation</td>
<td>Pahalaa</td>
<td>पहाला</td>
<td>पहाला</td>
</tr>
<tr>
<td>Short words</td>
<td>yr</td>
<td>यर</td>
<td>यार</td>
</tr>
<tr>
<td>Erroneous words</td>
<td>chaaapp</td>
<td>चाअप</td>
<td>छाप</td>
</tr>
</tbody>
</table>
Conclusions

- Used several classification techniques for solving the problem of language identification and NERC.
- For transliteration we have used a modified joint source-channel model.
- Transliteration performance can be improved using spelling variation techniques.
- Our system showed one of the best performance if EQMF ALL performance metric is considered.
- 2nd if EQMF ALL (No transliteration) is considered.