LIGA and Syllabification Approach for Language Identification and Back Transliteration:
Shared Task Report by DA-IICT

Shraddha Patel
Vaibhavi Desai
Problem Statement

Subtask 1: Query Word Labeling

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 (Hindi / Gujarati). The task is to label the words as E or L depending on whether it an English word, or a transliterated L-language word. And then, for each transliterated word, provide the correct transliteration in the native script (i.e., the script which is used for writing L).

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>palak paneer recipe</td>
<td>palak\H=पालक paneer\H=पनीर recipe\E</td>
</tr>
<tr>
<td>Maro phone bagadi gayo</td>
<td>Maro\G=મારો phone\E bagadi\G=మారో phone\E bagadi\G=పగడి gayo\G=ગయో</td>
</tr>
</tbody>
</table>
Language Identification Graph Approach (LIGA): [ Gujarati and Hindi ] - Constructing the Graph

- Constructing bi grams and tri grams of the words in the training data
- For each word in the training set, construct a simple graph and compute path matching scores for both languages using LIGA

Example:

LIGA Approach for training data
Calculating node and edge scores (tri-gram) for a set of three words “apple”, “apply” and “applied”
Language Identification Graph Approach (LIGA):
[ Gujarati and Hindi ] - Path Matching Scores

If the test word is “applies”, a LIGA graph can be constructed which will produce the following simple path:

app -> ppl -> pli -> lie -> ies

Calculating the Path Matching (PM) score for a language having training LIGA graph shown in figure 1 can be done as follows:

(adding all weights)
Total no of vertices = 11
Total no of edges = 8

PM = \( \frac{3}{11} + \frac{3}{11} + \frac{1}{11} + \frac{1}{11} + 0 + \frac{3}{8} + \frac{1}{8} + \frac{1}{8} + 0 \)
PM = 1.352
Language Identification Graph Approach (LIGA):
[ Gujarati and Hindi ] - Predicting the language

• For a language L, we calculate the path matching (PM) score for each word by constructing its bi-grams and tri-grams
• For each word of query set, the same method is applied to calculate the PM scores.

• A word in the query is labeled as “L” (Hindi, Gujarati or English) depending upon the maximum path matching score of that respective language.

Example:
• word in the query : “applies”
• PM score (English) = 1.352
• PM score (Hindi) = 0.112
• Hence, the word belongs to English and is labeled as “/E”
## LIGA : Results (Labelling Accuracy)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Hindi</th>
<th>English</th>
<th>Gujarati</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>77.3</td>
<td>71.0</td>
<td>97.6</td>
<td>16.7</td>
</tr>
<tr>
<td>Recall</td>
<td>78.2</td>
<td>79.5</td>
<td>98.6</td>
<td>25.0</td>
</tr>
<tr>
<td>F - score</td>
<td>78.8</td>
<td>75.0</td>
<td>98.1</td>
<td>20.0</td>
</tr>
<tr>
<td>Labeling Accuracy</td>
<td>77.1</td>
<td></td>
<td></td>
<td>96.3</td>
</tr>
</tbody>
</table>
LIGA : Error Analysis and Drawbacks

- Results highly depend on the size and credibility of the data sets.
- Single lettered words - not classified correctly
  - eg. “a”, “o”
- Problem with classification of proper nouns
  - eg. “Satyam”, “Delhi”
- Classification of words of different languages having same transliterations.
  - eg. Deep (both English and Hindi - दीप)
Back-transliteration : Rule Based Syllabification

Make syllables from words on the nearest consonants with at least one vowel. The last set of consonants can be taken as it is.

- Eg. Sudarshan = Su+da+rsha+n
- Eg. Vijay = Vi+ja+y
- Eg. Gada = Ga+da

- If the word ends in a vowel, the last syllable is constructed till the last vowel.
- Eg. Gada= Ga+ da (Instead of taking the last “a” as a separate syllable, append it with “d” and the last syllable thus becomes “da”)
Back-transliteration : Syllable Mapping

- Language of transliteration: P
- Language (Real): L
- Each syllable is then fed into a mapper where it gets mapped to a syllable in the language L. Some letters are mapped directly while some are mapped in combinations.

For instance, consider the Hindi word : khoobsoorat (खूबसूरत)
- Syllables: khoo, bsoo, ra, t
- mapping : ‘khoo’ : Since, ‘kh’ is mapped to the letter ‘ख’ instead of ‘k’ and ‘h’ individually mapped to their corresponding letters.
  - ‘oo’ is then mapped to ‘ऊ’

For mapping, a hash table is made where, each letter or combination of letters in P are mapped to one letter or letters in L.
- Such back transliterated syllables are then appended to form a complete word in language L.
Back-transliteration : Mapping to Dictionary

- S: naive word formed after syllable mapping.
- After constructing the naive word, the word is then looked for words in the dictionary of language L.

- If S maps directly to a word in the dictionary, it is taken as the output of the process.

- Else : Mapping for words: Mapping is done on a letter to letter basis in S.
- P: Word in the dictionary.
- Rules for mapping:
  - For each letter in S, the corresponding letter is taken in P. If the letters match, the check is continued. If the letters do not match, the alternate letter set of the letter in P is checked for. If the letter matches to any letter in the alternate letter set, the check is continued.
  - Alternate letter set: Some letters may have same phonetic representation or transliterated representations. For instance, Hindi letters "उ" and "ऊ" may be written as ‘u’ in English.
Back-transliteration : Mapping to dictionary and score calculation.

- Hence, with this process the search is narrowed to certain words where the check is done successfully.
- For example, the word ‘manav’ maps to
  - माणव
  - मानव

Score Calculation:
- A letter by letter comparison with the naive word is done. For every letter match an increment is given to the score. For every letter matching in the alternate letter set, 3/4th increment is given to the score. The word with the highest score, is the output.
### Back-transliteration: Results

<table>
<thead>
<tr>
<th></th>
<th>Hindi</th>
<th>Gujarati</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>9.6</td>
<td>46.4</td>
</tr>
<tr>
<td>Recall</td>
<td>52.3</td>
<td>46.2</td>
</tr>
<tr>
<td>F-Score</td>
<td>16.3</td>
<td>46.3</td>
</tr>
</tbody>
</table>
Back-transliteration: Error Analysis and Drawbacks

- Erroneous transliterations: The system does not give proper output for highly erroneous transliterations.

- Words having different phonetic representations but same transliterated representations may not be back transliterated efficiently.
  - eg. लाई and लायी
Acknowledgements

• Prasenjit Majumdar, DAIICT.
• Abhishek Shah, DAIICT.
• Monojit Chaudhry, Microsoft Research Lab, India
• Gokul Chittranjan, Microsoft Research Lab, India
Thank You