Shared Task – Machine Transliteration

Transliteration is defined as phonetic translation of names across languages. Transliteration of Named Entities (NEs) is necessary in many applications, such as machine translation, corpus alignment, cross-language IR, information extraction and automatic lexicon acquisition. All such systems call for high-performance transliteration, which is the focus of the shared task in the NEWS 2009 workshop. The objective of the shared task is to promote machine transliteration research by providing a common benchmarking platform for the community to evaluate the state-of-the-art technologies.

Task Description

The task is to develop machine transliteration system in one or more of the specified language pairs being considered for the task. Each language pair consists of a source and a target language. The training and development data sets released for each language pair are to be used for developing a transliteration system in whatever way that the participants find appropriate. At the evaluation time, a test set of names would be released, on which the participants are expected to produce for a ranked list of transliteration candidates in another language (ie. \( n \)-best transliterations), and this will be evaluated using common metrics. For every language pair the participants must submit one run that uses only the data provided by the NEWS workshop organizers in a given language pair (designated as “standard” runs), and this run would be used for common evaluation. Users may submit more runs (“non-standard”) each language pair that uses other data than those provided by the NEWS workshop; such runs would be evaluated and reported separately, but not as a part of common evaluation runs.

Important Dates

<table>
<thead>
<tr>
<th>Research Paper Submissions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Paper Submission Deadline</td>
<td>1-May-2009</td>
</tr>
</tbody>
</table>

| Shared Task                                     |
|------------------------------------------------|----------|
| Registration Opens                              | 16-Feb-2009|
| Registration Closes                             | 9-Apr-2009|
| Release Training/Development Data               | 16-Feb-2009|
| Release Test Data                               | 10-Apr-2009|
| Results Submission Due                          | 14-Apr-2009|
| Results Announcement                            | 29-Apr-2009|
| Task (short) Papers Due                         | 3-May-2009|

| For All Submissions                             |
|------------------------------------------------|----------|
| Acceptance Notification                         | 1-Jun-2009|
| Camera-Ready Copy Deadline                     | 7-Jun-2009|
| Workshop Date                                  | 7-Aug-2009|
Participation

1. Registration (16 Feb 2009)
   a. NEWS Shared Task opens for registration.
   b. Prospective participants are to register to the NEWS Workshop homepage.

2. Training & Development Data (16 Feb 2009)
   a. Registered participants are to obtain training and development data from the Shared Task organizer and/or the designated copyright owners of databases.

3. Evaluation Script (16 Mar 2009)
   a. A sample test set and expected user output format are to be released.
   b. An evaluation script which runs on the above two is to be released.
   c. The participants must make sure that their output is produced in a way that the evaluation script may run and produce the expected output.
   d. The same script (with held out test data and the user outputs) would be used for final evaluation.

4. Testing data (10 April 2009)
   a. The test data would be released on 10 Apr 2009, and the participants have a maximum of 4 days to submit their results in the expected format.
   b. Minimum of 1 standard run is required from every group on a given language pair; more (<= 5) may be submitted with appropriately tagged.
   c. Any runs that are non-standard must be tagged as such.

5. Results (29 April 2009)
   a. On 29 April 2009, the evaluation results would be announced and will be made available on the Workshop website.
   b. Note that only the scores (in respective metrics) of the participating systems on each language pairs would be published, and no explicit ranking of the participating systems would be published.
   c. Note that this is a shared evaluation task and not a competition; the results are meant to be used to evaluate systems on common data set with common metrics, and not to rank the participating systems. While the participants can cite the performance of their systems (scores on metrics) from the workshop report, they should not use any ranking information in their publications.
   d. Further, all participants should agree not to reveal identities of other participants in any of their publications unless you get permission from the other respective participants. If the participants want to remain anonymous in published results, they should inform the organizers (mzhang@i2r.a-star.edu.sg, a.kumaran@microsoft.com), at the time of registration. Note that the results of their systems would still be published, but with the participant identities masked. As a result, in this case, your organization name will still appear in the web site as one of participants, but it is not linked explicitly with your results.

6. Short Papers on Task (3 May 2009)
   a. Each submitting site is required to submit a 4-page system paper (short paper) for its submissions, including their approach, data used and the results.
   b. All system short papers will be included in the proceedings. Selected short papers will be presented in the NEWS 2009 workshop. Acceptance of the system shortpapers would be announced on 1 June 2009 together with that of other papers.
   c. All paper submission and review will be managed electronically through http://www.softconf.com.
**Languages Involved**

The tasks are to transliterate personal names or place names from a source to a target language as summarized in the following table.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language Family</th>
<th>Target</th>
<th>Language Family</th>
<th>Data Owner</th>
<th>No. of Entries</th>
<th>Task ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Chinese</td>
<td>Sino-Tibetan</td>
<td>I²R</td>
<td>&gt;30K</td>
<td>EnCh</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Japanese Katakana</td>
<td>Japonic, Altaic</td>
<td>CJK Institute²</td>
<td>&gt;25K</td>
<td>EnJa</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Korean Hangul</td>
<td>Altaic</td>
<td>CJK Institute²</td>
<td>&gt;7K</td>
<td>EnKo</td>
</tr>
<tr>
<td>Japanese Name (in English)</td>
<td>Indo-European</td>
<td>Japanese Kanji</td>
<td>Japonic, Altaic</td>
<td>CJK Institute²</td>
<td>&gt;20K</td>
<td>JnKo</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Hindi</td>
<td>Indo-European</td>
<td>MSR India</td>
<td>~15K</td>
<td>EnHi</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Tamil</td>
<td>Dravidian</td>
<td>MSR India</td>
<td>~15K</td>
<td>EnTa</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Kannada</td>
<td>Dravidian</td>
<td>MSR India</td>
<td>~15K</td>
<td>EnKa</td>
</tr>
<tr>
<td>English</td>
<td>Indo-European</td>
<td>Russian</td>
<td>Slavic</td>
<td>MSR India</td>
<td>~10K</td>
<td>EnRu</td>
</tr>
</tbody>
</table>

Notes:
The names given in the training sets for Chinese, Japanese and Korean languages are Western names and their CJK transliterations; the Japanese Name (in English) -> Japanese Kanji data set consists only of native Japanese names. The Indic data set (Hindi, Tamil, Kannada) consists of a mix of Indian and Western names.

**SAMPLES:**
- English -> Chinese: Timothy -> 蒂莫西
- English -> Japanese Katakana: Harrington -> ハリントン
- English -> Korean Hangul: Bennett -> 베넷
- Japanese Name in English -> Japanese Kanji: Akihiro -> 秋宏
- English -> Hindi: San Francisco -> सैन फ्रांसिस्को
- English -> Tamil: London -> லத்தொன்
- English -> Kannada: Tokyo -> ಟೊಕ್ಯೋ
- English -> Russian: Moscow -> Москва

**Standard Databases**

<table>
<thead>
<tr>
<th>Data</th>
<th>Size</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Data (Parallel):</strong></td>
<td></td>
<td>Training Data is used for training a basic transliteration system.</td>
</tr>
<tr>
<td>Paired names between source and target languages</td>
<td>5K – 40K</td>
<td></td>
</tr>
<tr>
<td><strong>Development Data (Parallel):</strong></td>
<td></td>
<td>Development Data is in addition to the Training data, which is used for system fine-tuning of parameters in case of need. Participants are</td>
</tr>
</tbody>
</table>

¹ There may be a nominal fee for these data sets. Please contact the data owner for obtaining the databases.
1. Participants will need to obtain licenses from the respective copyright owners. NEWS will provide the contact details of each individual database. The data would be provided in Unicode UTF-8 encoding, in XML format; the results are expected to be submitted in XML format. The XML formats will be announced at the workshop website.

2. The data are provided in 3 sets as described in the table above.

3. Name pairs are distributed as-is, as provided by the respective creators.
   a. While the databases are mostly manually checked, there may be still inconsistency (that is, non-standard usage, region-specific usage, errors, etc.) or incompleteness (that is, not all right variations may be covered).
   b. The participants may use any method to further clean up the data provided.
      i. If they are cleaned up manually, we appeal that such data be provided back to the organizers for redistribution to all the participating groups in that language pair; such sharing benefits all participants, and further ensures that the evaluation provides normalization with respect to data quality.
      ii. If automatic cleanup were used, such cleanup would be considered a part of the system fielded, and hence not required to be shared with all participants.

4. We expect that the participants to use only the data (parallel names) provided by the Shared Task for transliteration task for a standard run to ensure a fair evaluation. One such run (using only the data provided by the shared task) is mandatory for all participants for a given language pair that they participate in.

5. If more data (either parallel names data or monolingual data) were used, then all such runs using extra data must be marked as non-standard. For such non-standard runs, it is required to disclose the size and characteristics of the data used in the system paper.

6. A participant may submit a maximum of 5 runs for a given language pair (including the mandatory “standard” run)

**Paper Format**

Paper submissions to NEWS 2009 should follow the ACL-IJCNLP-2009 paper submission policy, including paper format, blind review policy and title and author format convention. Full papers (research paper) are in two-column format without exceeding eight (8) pages of content plus one extra page for references and short papers (task paper) are also in two-column format without exceeding four (4) pages, including references. Submission must conform to the official ACL-IJCNLP-2009 style guidelines. For details, please refer to [http://www.acl-ijcnlp-2009.org/main/authors/stylefiles/index.html](http://www.acl-ijcnlp-2009.org/main/authors/stylefiles/index.html).

**Evaluation Metrics**

We plan to measure the quality of the transliteration task using the following 5 metrics. We accept up to 10 output candidates in a ranked list for each input entry.
Since a given source name may have multiple correct target transliterations, all these alternatives are treated equally in the evaluation. That is, any of these alternatives are considered as a correct transliteration, and the first correct transliteration in the ranked list is accepted as a correct hit.

The following notation is assumed further:

- $N$: Total number of names (source words) in test set
- $n_i$: Number of reference transliterations for $i$-th name in the test set ($n_i \geq 1$)
- $r_{i,j}$: $j$-th reference transliteration for $i$-th name in the test set
- $c_{i,k}$: $k$-th candidate transliteration (system output) for $i$-th name in the test set ($1 \leq k \leq 10$).
- $K_i$: Number of candidate transliterations produced by a transliteration system

1. **Word Accuracy in Top-1 (ACC)**
   Also known as Word Error Rate, it measures correctness of the first transliteration candidate in the n-best candidate list produced by a transliteration system. $ACC = 1$ means that all top candidates are correct transliterations i.e. they match one of the references, and $ACC = 0$ means that none of the top candidates are correct.

   $$ACC = \frac{1}{N} \sum_{i=1}^{N} \{1 \text{ if } \exists r_{i,j} : r_{i,j} = c_{i,1}; 0 \text{ otherwise}\}$$

2. **Fuzziness in Top-1 (Mean F-score)**
   The mean F-score measures how different, on average, the top transliteration candidate is from its closest reference. F-score for each source word is a function of Precision and Recall and equals 1 when the top candidate matches one of the references, and 0 when there are no common characters between the candidate and any of the references.

   Precision and Recall are calculated based on the length of the Longest Common Subsequence between a candidate and a reference:

   $$LCS(c, r) = \frac{1}{2}(\text{length}(c) + \text{length}(r) - ED(c, r))$$

   where $ED(c, r)$ is the edit distance. For example, the longest common subsequence between “abcd” and “afcde” is “acd” and its length is 3. The best matching reference, that is, the reference for which the edit distance has the minimum is taken for calculation. If the best matching reference $r_{i,m}$ is given by

   $$r_{i,m} = \arg\min_j(ED(c_{i,1}, r_{i,j}))$$

   Recall, Precision and F-score for $i$-th word are then calculated as

   $$R_i = \frac{LCS(c_{i,1}, r_{i,m})}{\text{length}(r_{i,m})} \quad P_i = \frac{LCS(c_{i,1}, r_{i,m})}{\text{length}(c_{i,1})} \quad F_i = \frac{2R_i \times P_i}{R_i + P_i}$$

   a. The length is computed in distinct Unicode characters.
   b. No distinction is made on different character types of a language (e.g., vowel vs. consonants vs. combining diereses’ etc.)

3. **Mean Reciprocal Rank (MRR)**
   Measures traditional MRR *for any right answer* produced by the system, from among $n_i$ candidates. $1$/MRR tells approximately the average rank of the correct transliteration. MRR closer to 1 implies that the correct answer is mostly produced close to the top of the n-best lists.
If a candidate that matches one of the references is at the $j$-th position in the $n$-best list, its rank equals $j$ and its reciprocal rank equals $1/j$. When none of the candidates matches any of the references, the reciprocal rank of the “matching” candidate is 0.

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \left\{ \min_{j} \frac{1}{j} \text{ if } \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k}; 0 \text{ otherwise} \right\}$$

Example of MRR calculation: If reference translliterations of a source word are “abc”, “xyz”, and the $n$-best list of candidates generated by a transliteration system for the source word is “abx”, “xyz”, “xyr”, “xyz”, “abf”, “abc”, … (at most 10 candidates) starting from the most likely one, the first candidate that matches one of the references is “xyz” at position 3 in the $n$-best list. Therefore, the contribution of this source word transliteration to MRR is $(1/3)/N$. Although there is another matching candidate in the $n$-best list (“abc” at position 5), only the topmost candidate matching any reference is taken into account.

4. $MAP_{ref}$

Measures tightly the precision in the $n$-best candidates for $i$-th source name, for which $n_i$ reference translliterations are available. If all of the references are produced, then the MAP is 1.

$$MAP_{ref} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \left( \sum_{k=1}^{n_i} \frac{\text{number of correct candidates for } i-\text{th word; in } k-\text{best}}{k} \right)$$

5. $MAP_{10}$

Measures the precision in the 10-best candidates for $i$-th source name provided by the candidate system. In general, the higher $MAP_{10}$ is, the better is the quality of the transliteration system in capturing the multiple references. Note that the number of reference translliterations may be more or less than 10. If the number of reference translliterations is below 10, then $MAP_{10}$ can never be equal to 1. Only if the number of reference translliterations for every source word is at least 10, then $MAP_{10}$ could possibly be equal to 1.

$$MAP_{10} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{10} \left( \sum_{k=1}^{10} \frac{\text{number of correct candidates for } i-\text{th word; in } k-\text{best}}{k} \right)$$

Note that in general $MAP_m$ measures the “goodness in m-best” candidate list. We use $m=10$ because we have asked the systems to produce up to 10 candidates for every source name in the test set.

6. $MAP_{sys}$

Measures the precision in the top $K_i$-best candidates produced by the system for $i$-th source name, for which $n_i$ reference translliterations are available. This measure allows the systems to produce variable number of translliterations, based on their confidence in identifying and producing correct translliterations. If all of the $n_i$ references are produced in the top $n_i$ candidates (that is, $K_i = n_i$, and all of them are correct), then the MAP is 1.

$$MAP_{sys} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K_i} \left( \sum_{k=1}^{K_i} \frac{\text{number of correct candidates for } i-\text{th word; in } k-\text{best}}{k} \right)$$
Contact Us:
If you have any questions about this share task and the database, please email to

Dr. Haizhou Li
Institute for Infocomm Research (I2R), A*STAR
1 Fusionopolis Way
#08-05 South Tower, Connexis
Singapore 138632
hli@i2r.a-star.edu.sg

Dr. A. Kumaran
Microsoft Research India
Scientia, 196/36, Sadashivnagar 2nd Main Road
Bangalore 560080 INDIA
a.kumaran@microsoft.com

Mr. Jack Halpern
The CJK Dictionary Institute (CJK Data)
Komine Building (3rd & 4th floors)
34-14, 2-chome, Tohoku, Niiza-shi
Saitama 352-0001 JAPAN
jack@cjki.org
Appendix A: Training/Dev Data

- File Naming Conventions:
  - NEWS09_train_XXYY_nnnn.xml,
    NEWS09_dev_XXYY_nnnn.xml
    NEWS09_test_XXYY_nnnn.xml
    - XX: Source Language
    - YY: Target Language
    - nnnn: size of parallel/monolingual names ("25K", "10000", etc.)

- File Formats:
  - All data would be made available in XML formats (Appendix A).

- Data Encoding Formats:
  - The data would be in Unicode, in UTF-8 encoding. The results are expected to be submitted in UTF-8 format only, and in the XML format specified.

File: NEWS2009_Train_EnHi_25K.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TransliterationCorpus
    CorpusID = "NEWS2009-Train-EnHi-25K"
    SourceLang = "English"
    TargetLang = "Hindi"
    CorpusType = "Train|Dev"
    CorpusSize = "25000"
    CorpusFormat = "UTF8">
    <Name ID="1">
        <SourceName>eeeeee1</SourceName>
        <TargetName ID="1">hhhhhh1_1</TargetName>
        <TargetName ID="2">hhhhhh1_2</TargetName>
        ...
        <TargetName ID="n">hhhhhh1_n</TargetName>
    </Name>
    <Name ID="2">
        <SourceName>eeeeee2</SourceName>
        <TargetName ID="1">hhhhhh2_1</TargetName>
        <TargetName ID="2">hhhhhh2_2</TargetName>
        ...
        <TargetName ID="m">hhhhhh2_m</TargetName>
    </Name>
    ...
    <!-- rest of the names to follow -->
    ...
</TransliterationCorpus>```
**Appendix B: Submission of Results**

- **File Naming Conventions:**
  - NEWS09_Result_XXYY_gggg_nn_description.xml
    - XX: Source
    - YY: Target
    - gggg: Group ID
    - nn: run ID. Note that run ID “1” stands for “standard” run where only the provided data are allowed to be used. Run ID “2-5” means “non-standard” run where additional data can be used.
    - description: Description of the run

- **File Formats:**
  - All results would be submitted in XML formats (Appendix B).

- **Data Encoding Formats:**
  - The data would be in Unicode, in UTF-8 encoding. The results are expected to be submitted in UTF-8 format only.

*Example: NEWS2009_EnHi_TUniv_01_StdRunHMMBased.xml*

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TransliterationTaskResults
  SourceLang = "English"
  TargetLang = "Hindi"
  GroupID = "Trans University"
  RunID = "1"
  RunType = "Standard"
  Comments = "HMM Run with params: alpha=0.8 beta=1.25">
  <Name ID="1">
    <SourceName>eeeeee1</SourceName>
    <TargetName ID="1">hhhhhh11</TargetName>
    <TargetName ID="2">hhhhhh12</TargetName>
    <TargetName ID="3">hhhhhh13</TargetName>
    ...
    <TargetName ID="10">hhhhhh110</TargetName>
  </Name>
  <!-- Participants to provide their top 10 candidate transliterations -->

  <Name ID="2">
    <SourceName>eeeeee2</SourceName>
    <TargetName ID="1">hhhhhh21</TargetName>
    <TargetName ID="2">hhhhhh22</TargetName>
    <TargetName ID="3">hhhhhh23</TargetName>
    ...
    <TargetName ID="10">hhhhhh110</TargetName>
  </Name>
  <!-- Participants to provide their top 10 candidate transliterations -->

  <!-- All names in test corpus to follow -->
  ...
</TransliterationTaskResults>
```