Whitepaper of NEWS 2012 Shared Task on Machine Transliteration

Min Zhang†, Haizhou Li†, Ming Liu†, A Kumaran‡

†Institute for Infocomm Research, A*STAR, Singapore 138632
{mzhang,hli,mliu}@i2r.a-star.edu.sg
‡Multilingual Systems Research, Microsoft Research India
A.Kumaran@microsoft.com

Abstract

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 shared task in the NEWS 2012 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.

1 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 source names only would be released, on which the participants are expected to produce a ranked list of transliteration candidates in another language (i.e. n-best transliterations), and this will be evaluated using common metrics. For every language pair the participants must submit at least one run that uses only the data provided by the NEWS workshop organisers in a given language pair (designated as “standard” run, primary submission). Users may submit more “standard” runs. They may also submit several “non-standard” runs for each language pair that use other data than those provided by the NEWS 2012 workshop; such runs would be evaluated and reported separately.

2 Important Dates

<table>
<thead>
<tr>
<th>Research paper submission deadline</th>
<th>25 March 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared task</td>
<td></td>
</tr>
<tr>
<td>Registration opens</td>
<td>18 Jan 2012</td>
</tr>
<tr>
<td>Registration closes</td>
<td>11 Mar 2012</td>
</tr>
<tr>
<td>Training/Development data release</td>
<td>20 Jan 2012</td>
</tr>
<tr>
<td>Test data release</td>
<td>12 Mar 2012</td>
</tr>
<tr>
<td>Results Submission Due</td>
<td>16 Mar 2012</td>
</tr>
<tr>
<td>Results Announcement</td>
<td>20 Mar 2012</td>
</tr>
<tr>
<td>Task (short) Papers Due</td>
<td>25 Mar 2012</td>
</tr>
<tr>
<td>For all submissions</td>
<td></td>
</tr>
<tr>
<td>Acceptance Notification</td>
<td>20 April 2012</td>
</tr>
<tr>
<td>Camera-Ready Copy Deadline</td>
<td>30 April 2012</td>
</tr>
<tr>
<td>Workshop Date</td>
<td>12/13/14 July 2012</td>
</tr>
</tbody>
</table>

3 Participation

1. Registration (18 Jan 2012)
   (a) NEWS Shared Task opens for registration.
   (b) Prospective participants are to register to the NEWS Workshop homepage.

2. Training & Development Data (20 Jan 2012)
   (a) Registered participants are to obtain training and development data from the Shared Task organiser and/or the designated copyright owners of databases.
   (b) All registered participants are required to participate in the evaluation of at least one language pair, submit the results and a short paper and attend the workshop at ACL 2012.

3. Test data (12 March 2012)
The test data would be released on 12 March 2012, and the participants have a maximum of 5 days to submit their results in the expected format.

One “standard” run must be submitted from every group on a given language pair. Additional “standard” runs may be submitted, up to 4 “standard” runs in total. However, the participants must indicate one of the submitted “standard” runs as the “primary submission”. The primary submission will be used for the performance summary. In addition to the “standard” runs, more “non-standard” runs may be submitted. In total, maximum 8 runs (up to 4 “standard” runs plus up to 4 “non-standard” runs) can be submitted from each group on a registered language pair. The definition of “standard” and “non-standard” runs is in Section 5.

Any runs that are “non-standard” must be tagged as such.

The test set is a list of names in source language only. Every group will produce and submit a ranked list of transliteration candidates in another language for each given name in the test set. Please note that this shared task is a “transliteration generation” task, i.e., given a name in a source language one is supposed to generate one or more transliterations in a target language. It is not the task of “transliteration discovery”, i.e., given a name in the source language and a set of names in the target language evaluate how to find the appropriate names from the target set that are transliterations of the given source name.

4. Results (20 March 2012)

On 20 March 2012, the evaluation results would be announced and will be made available on the Workshop website.

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.

5. Short Papers on Task (25 March 2012)

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 on either test set or development set or by n-fold cross validation on training set.

The review of the system papers will be done to improve paper quality and readability and make sure the authors’ ideas and methods can be understood by the workshop participants. We are aiming at accepting all system papers, and selected ones will be presented orally in the NEWS 2012 workshop.

All registered participants are required to register and attend the workshop to introduce your work.

All paper submission and review will be managed electronically through https://www.softconf.com/acl2012/news2012/.
4 Language Pairs

The tasks are to transliterate personal names or place names from a source to a target language as summarised in Table 1. NEWS 2012 Shared Task offers 14 evaluation subtasks, among them ChEn and ThEn are the back-transliteration of EnCh and EnTh tasks respectively. NEWS 2012 releases training, development and testing data for each of the language pairs. NEWS 2012 continues all language pairs that were evaluated in NEWS 2011. In such cases, the training and development data in the release of NEWS 2012 are the same as those in NEWS 2011. However, the test data in NEWS 2012 are entirely new.

Please note that in order to have an accurate study of the research progress of machine translation technology, different from previous practice, the test/reference sets of NEWS 2011 are not released to the research community. Instead, we use the test sets of NEWS 2011 as progress test sets in NEWS 2012. NEWS 2012 participants are requested to submit results on the NEWS 2012 progress test sets (i.e., NEWS 2011 test sets). By doing so, we would like to do comparison studies by comparing the NEWS 2012 and NEWS 2011 results on the progress test sets. We hope that we can have some insightful research findings in the progress studies.

The names given in the training sets for Chinese, Japanese, Korean, Thai and Persian languages are Western names and their respective transliterations; the Japanese Name (in English) → Japanese Kanji data set consists only of native Japanese names; the Arabic data set consists only of native Arabic names. The Indic data set (Hindi, Tamil, Kannada, Bangla) consists of a mix of Indian and Western names.

Examples of transliteration:

**English → Chinese**
Timothy → 蒂莫西

**English → Japanese Katakana**
Harrington → ハリントン

**English → Korean Hangul**
Bennett → 베테넷

**Japanese name in English → Japanese Kanji**
Akihiro → 秋宏

**English → Hindi**
San Francisco → सैन फ्रांसिस्को

**English → Tamil**
London → லங்கோணம்

**English → Kannada**
Tokyo → ಟೊಕ್ಯೋ

**Arabic → Arabic name in English**
خالد → Khalid

5 Standard Databases

**Training Data (Parallel)**
Paired names between source and target languages; size 7K – 37K.
Training Data is used for training a basic transliteration system.

**Development Data (Parallel)**
Paired names between source and target languages; size 1K – 2.8K.
Development Data is in addition to the Training data, which is used for system fine-tuning of parameters in case of need. Participants are allowed to use it as part of training data.

**Testing Data**
Source names only; size 1K – 2K.
This is a held-out set, which would be used for evaluating the quality of the transliterations.

**Progress Testing Data**
Source names only; size 0.6K – 2.6K.
This is the NEWS 2011 test set, it is held-out for progress study.

1. Participants will need to obtain licenses from the respective copyright owners and/or agree to the terms and conditions of use that are given on the downloading website (Li et al., 2004; MSRI, 2010; CJKI, 2010). NEWS 2011 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 UTF-8 encoding in XML format. The XML formats details are available in Appendix A.

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

3. Name pairs are distributed as-is, as provided by the respective creators.
### Table 1: Source and target languages for the shared task on transliteration.

<table>
<thead>
<tr>
<th>Name origin</th>
<th>Source script</th>
<th>Target script</th>
<th>Data Owner</th>
<th>Data Size</th>
<th>Task ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western</td>
<td>English</td>
<td>Chinese</td>
<td>Institute for Infocomm Research</td>
<td>37K 2.8K 2K 1K</td>
<td>EnCh</td>
</tr>
<tr>
<td>Western</td>
<td>Chinese</td>
<td>English</td>
<td>Institute for Infocomm Research</td>
<td>28K 2.7K 2.2K 1K</td>
<td>ChEn</td>
</tr>
<tr>
<td>Western</td>
<td>English</td>
<td>Korean Hangul</td>
<td>CJK Institute</td>
<td>7K 1K 609 1K</td>
<td>EnKo</td>
</tr>
<tr>
<td>Western</td>
<td>English</td>
<td>Japanese Katakana</td>
<td>CJK Institute</td>
<td>26K 2K 1.8K 1K</td>
<td>EnJa</td>
</tr>
<tr>
<td>Japanese</td>
<td>English</td>
<td>Japanese Kanji</td>
<td>CJK Institute</td>
<td>10K 2K 571 1K</td>
<td>JnJk</td>
</tr>
<tr>
<td>Arabic</td>
<td>Arabic</td>
<td>English</td>
<td>CJK Institute</td>
<td>27K 2.5K 2.6K 1K</td>
<td>ArEn</td>
</tr>
<tr>
<td>Mixed</td>
<td>English</td>
<td>Hindi</td>
<td>Microsoft Research India</td>
<td>12K 1K 1K 1K</td>
<td>EnHi</td>
</tr>
<tr>
<td>Mixed</td>
<td>English</td>
<td>Tamil</td>
<td>Microsoft Research India</td>
<td>10K 1K 1K 1K</td>
<td>EnTa</td>
</tr>
<tr>
<td>Mixed</td>
<td>English</td>
<td>Kannada</td>
<td>Microsoft Research India</td>
<td>10K 1K 1K 1K</td>
<td>EnKa</td>
</tr>
<tr>
<td>Mixed</td>
<td>English</td>
<td>Bangla</td>
<td>Microsoft Research India</td>
<td>13K 1K 1K 1K</td>
<td>EnBa</td>
</tr>
<tr>
<td>Western</td>
<td>English</td>
<td>Thai</td>
<td>NECTEC</td>
<td>27K 2K 2K 1K</td>
<td>EnTh</td>
</tr>
<tr>
<td>Western</td>
<td>Thai</td>
<td>English</td>
<td>NECTEC</td>
<td>25K 2K 1.9K 1K</td>
<td>ThEn</td>
</tr>
<tr>
<td>Western</td>
<td>English</td>
<td>Persian</td>
<td>Sarvnaz Karimi / RMIT</td>
<td>10K 2K 2K 1K</td>
<td>EnPe</td>
</tr>
<tr>
<td>Western</td>
<td>English</td>
<td>Hebrew</td>
<td>Microsoft Research India</td>
<td>9.5K 1K 1K 1K</td>
<td>EnHe</td>
</tr>
</tbody>
</table>

(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 organisers for redistribution to all the participating groups in that language pair; such sharing benefits all participants, and further ensures that the evaluation provides normalisation 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. **Standard Runs** 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. **Non-standard Runs** 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 8 runs for a given language pair (including the mandatory 1 “standard” run marked as “primary submission”).

6. **Paper Format**

Paper submissions to NEWS 2012 should follow the ACL 2012 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 two (2) extra page for references and short papers (task paper) are also in two-column format without exceeding four (4) pages content plus two (2) extra page for references. Submission must conform to the official ACL 2012 style guidelines. For details, please refer to the ACL 2012 website.

7. **Evaluation Metrics**

We plan to measure the quality of the transliteration task using the following 4 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.

---

2http://www.ACL2012.org/
The following notation is further assumed:

- \( N \) : Total number of names (source words) in the 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 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.

   \[
   ACC = \frac{1}{N} \sum_{i=1}^{N} \left\{ \begin{array}{ll}
   1 & \text{if } \exists r_{i,j} : r_{i,j} = c_{i,1} \\
   0 & \text{otherwise}
   \end{array} \right. 
   \]
   \hspace{1cm} (1)

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:

   \[
   LCS(c, r) = \frac{1}{2} (|c| + |r| - ED(c, r)) \hspace{1cm} (2)
   \]

   where \( ED \) is the edit distance and \( |x| \) is the length of \( x \). 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 is given by

   \[
   r_{i,m} = \arg \min_j (ED(c_{i,1}, r_{i,j})) \hspace{1cm} (3)
   \]

   then Recall, Precision and F-score for \( i \)-th word are calculated as

   \[
   R_i = \frac{LCS(c_{i,1}, r_{i,m})}{|r_{i,m}|} \hspace{1cm} (4)
   \]
   \[
   P_i = \frac{LCS(c_{i,1}, r_{i,m})}{|c_{i,1}|} \hspace{1cm} (5)
   \]
   \[
   F_i = \frac{2R_i P_i}{R_i + P_i} \hspace{1cm} (6)
   \]

- The length is computed in distinct Unicode characters.
- 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 the 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.

   \[
   RR_i = \left\{ \begin{array}{ll}
   \min_j \frac{1}{j} & \text{if } \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k} \\
   0 & \text{otherwise}
   \end{array} \right. 
   \]
   \hspace{1cm} (7)

   \[
   MRR = \frac{1}{N} \sum_{i=1}^{N} RR_i \hspace{1cm} (8)
   \]

4. **MAP**

   Measures tightly the precision in the n-best candidates for \( i \)-th source name, for which reference transliterations are available. If all of the references are produced, then the MAP is 1. Let’s denote the number of correct candidates for the \( i \)-th source word in \( k \)-best list as \( num(i,k) \). MAP is then given by

   \[
   MAP_{ref} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \left( \sum_{k=1}^{n_i} num(i,k) \right) \hspace{1cm} (9)
   \]

8. **Contact Us**

If you have any questions about this share task and the database, please email to

**Mr. Ming Liu**

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

References

http://www.cjk.org/.

[Li et al.2004] Haizhou Li, Min Zhang, and Jian Su.

http://research.microsoft.com/india.
A Training/Development Data

- File Naming Conventions:
  NEWS12_train_XXYY_nnnn.xml
  NEWS12_dev_XXYY_nnnn.xml
  NEWS12_test_XXYY_nnnn.xml
  NEWS11_test_XXYY_nnnn.xml
  (progress test sets)
  - XX: Source Language
  - YY: Target Language
  - nnnn: size of parallel/monolingual names ("25K", "10000", etc)

- File formats:
  All data will be made available in XML formats (Figure 1).

- Data Encoding Formats:
  The data will be in Unicode UTF-8 encoding files without byte-order mark, and in the XML format specified.

B Submission of Results

- File Naming Conventions:
  You can give your files any name you like. During submission online you will need to indicate whether this submission belongs to a “standard” or “non-standard” run, and if it is a “standard” run, whether it is the primary submission.

- File formats:
  All data will be made available in XML formats (Figure 2).

- Data Encoding Formats:
  The results are expected to be submitted in UTF-8 encoded files without byte-order mark only, and in the XML format specified.
<?xml version="1.0" encoding="UTF-8"?>

<TransliterationCorpus
    CorpusID = "NEWS2012-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>

Figure 1: File: NEWS2012_Train_EnHi_25K.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>

Figure 2: Example file: NEWS2012_EnHi_TUniv_01.StdRunHMMBased.xml