Whitepaper of NEWS 2009 Machine Transliteration Shared Task*

Haizhou Li†, A Kumaran‡, Min Zhang† and Vladimir Pervouchine†

†Institute for Infocomm Research, A*STAR, Singapore 138632
{hli,mzhang,vpervouchine}@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 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.

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 one run that uses only the data provided by the NEWS workshop organisers in a given language pair (designated as "standard" runs). Users may submit more runs ("non-standard") for each language pair that uses other data than those provided by the NEWS 2009 workshop; such runs would be evaluated and reported separately.

*http://www.acl-ijcnlp-2009.org/workshops/NEWS2009/

2 Important Dates

Research paper submission deadline	1 May 2009
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

3 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 organiser 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.

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 19–26,
Suntec, Singapore, 7 August 2009.

(d) The same script (with held out test data and the user outputs) would be used for final evaluation.

4. Test 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) Only 1 "standard" run must be submitted from every group on a given language pair; more "non-standard" runs (0 to 4) may be submitted. In total, maximum 5 runs (1 "standard" run plus up to 4 "non-standard" runs) can be submitted from each group on a registered language pair.

   (c) Any runs that are "non-standard" must be tagged as such.

   (d) 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.

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

   (b) All system short papers will be included in the proceedings. Selected short papers will be presented orally in the NEWS 2009 workshop. Reviewers' comments for all system short papers and the acceptance notification for the system short papers for oral presentation would be announced on 1 June 2009 together with that of other papers.

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

   (d) All paper submission and review will be managed electronically through https://www.softconf.com/acl-ijcnlp09/NEWS/.

4 Languages Involved

The tasks are to transliterate personal names or place names from a source to a target language as summarised in Table 1.

Source language	Target language	Data Owner	Approx. Data Size	Task ID
English	Chinese	Institute for Infocomm Research	30K	EnCh
English	Japanese Katakana	CJK Institute	25K	EnJa
English	Korean Hangul	CJK Institute	7K	EnKo
Japanese name (in English)	Japanese Kanji	CJK Institute	20K	JnJk
English	Hindi	Microsoft Research India	15K	EnHi
English	Tamil	Microsoft Research India	15K	EnTa
English	Kannada	Microsoft Research India	15K	EnKa
English	Russian	Microsoft Research India	10K	EnRu

Table 1: Source and target languages for the shared task on transliteration.

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.

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 फ्रान्सस्कि को

English → Tamil
London → லண்டன்

English → Kannada
Tokyo → ಟೋಕ್ಯೋ

English → Russian
Moscow → Москва

5 Standard Databases

Training Data (Parallel)
Paired names between source and target languages; size 5K – 40K.
Training Data is used for training a basic transliteration system.

Development Data (Parallel)
Paired names between source and target languages; size 1K – 2K.
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 – 3K.
This is a held-out set, which would be used for evaluating the quality of the transliterations.

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; Kumaran and Kellner, 2007; MSRI, 2009; CJKI, 2009). NEWS 2009 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 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 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. 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 1 "standard" run).

6 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 the website[2].

7 Evaluation Metrics

We plan to measure the quality of the transliteration task using the following 6 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 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.

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{l} 1 \text{ if } \exists r_{i,j} : r_{i,j} = c_{i,1}; \\ 0 \text{ otherwise} \end{array}\right\} \quad (1)$$

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:

$$LCS(c, r) = \frac{1}{2}\left(|c| + |r| - ED(c, r)\right) \quad (2)$$

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

$$r_{i,m} = \arg\min_{j}\left(ED(c_{i,1}, r_{i,j})\right) \quad (3)$$

[2] http://www.acl-ijcnlp-2009.org/main/authors/stylefiles/index.html

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

$$R_i = \frac{LCS(c_{i,1}, r_{i,m})}{|r_{i,m}|} \quad (4)$$

$$P_i = \frac{LCS(c_{i,1}, r_{i,m})}{|c_{i,1}|} \quad (5)$$

$$F_i = 2\frac{R_i \times P_i}{R_i + P_i} \quad (6)$$

- 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.

$$RR_i = \left\{ \begin{array}{l} \min_j \frac{1}{j} \text{ if } \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k}; \\ 0 \text{ otherwise} \end{array} \right\} \quad (7)$$

$$MRR = \frac{1}{N}\sum_{i=1}^{N} RR_i \quad (8)$$

4. MAP$_{ref}$ 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$_{ref}$ is then given by

$$MAP_{ref} = \frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i} num(i,k)\right) \quad (9)$$

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 transliterations may be more or less than 10. If the number of reference transliterations is below 10, then MAP$_{10}$ can never be equal to 1. Only if the number of reference transliterations for every source word is at least 10, then MAP$_{10}$ could possibly be equal to 1.

$$MAP_{10} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{10}\left(\sum_{k=1}^{10} num(i,k)\right) \quad (10)$$

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 transliterations are available. This measure allows the systems to produce variable number of transliterations, based on their confidence in identifying and producing correct transliterations. 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$_{sys}$ is 1.

$$MAP_{sys} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i}\left(\sum_{k=1}^{K_i} num(i,k)\right) \quad (11)$$

8 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. Kurt Easterwood
The CJK Dictionary Institute (CJK Data)
Komine Building (3rd & 4th floors)
34-14, 2-chome, Tohoku, Niiza-shi
Saitama 352-0001 JAPAN
akurt@cjki.org

References

CJKI. 2009. CJK Institute. http://www.cjk.org/.

A Kumaran and T. Kellner. 2007. A generic framework for machine transliteration. In *Proc. SIGIR*, pages 721–722.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proc. 42nd ACL Annual Meeting*, pages 159–166, Barcelona, Spain.

MSRI. 2009. Microsoft Research India. http://research.microsoft.com/india.

Appendix A: Training/Development 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 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.

Appendix B: Submission of Results

- File Naming Conventions:
  `NEWS09_result_XXYY_gggg_nn_descr.xml`
  - `XX`: Source Language
  - `YY`: Target Language
  - `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.
  - `descr`: Description of the run.
- 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 = "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>
```

Figure 1: File: NEWS2009_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>

        <!-- Participants to provide their
        top 10 candidate transliterations -->
    </Name>
    <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>
        <!-- Participants to provide their
        top 10 candidate transliterations -->
    </Name>
    ...
    <!-- All names in test corpus to follow -->
    ...
</TransliterationTaskResults>
```

Figure 2: Example file: NEWS2009_EnHi_TUniv_01_StdRunHMMBased.xml