Whitepaper on NEWS 2018 Shared Task on Machine Transliteration

Nancy Chen†, Xiangyu Duan‡, Min Zhang‡, Rafael E. Banchs†, Haizhou Li⋆

†Institute for Infocomm Research, A*STAR, Singapore
{nfychen,rembanchs}@i2r.a-star.edu.sg

‡Soochow University, China 215006
{xiangyuduan,minzhang}@suda.edu.cn

⋆National University of Singapore, Singapore
haizhou.li@nus.edu.sg

Abstract

Transliteration is defined as the phonetic translation of names across languages. Transliteration of Named Entities (NEs) is a necessary subtask 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 NEWS 2018.

1 Task Description

The objective of the Shared Task on Named Entity Transliteration at NEWS 2018 is to promote machine transliteration research by providing a common benchmarking platform for the research community to evaluate state-of-the-art approaches to this problem. The task is to develop machine transliteration and/or back-transliteration systems in one or more of the provided language pairs.

For each language pair, training and development data sets containing source and target name pairs are released for participating teams to train their systems. At the evaluation time, test sets of source names only will be released, on which participants are expected to produce a ranked list of transliteration and/or back-transliteration candidates in the target language. The results will be automatically evaluated by using the same metrics used in previous editions of the shared task.

This year's shared task focuses mainly on "standard" submissions, i.e. output results from systems that have been trained only with the data provided by the shared task organizing team. This will ensure that all results for the same task are comparable across the different systems. Participants may submit several "standard" runs for each of the task they participate in. Those participants interested in submitting "non-standard" runs, i.e. output results from systems that use additional data during the training phase, still will be able to do so. However such runs will be evaluated and reported separately.

2 Important Dates

Train/Development data release	12 March 2018
Test data release	07 May 2018
Results Submission Due	14 May 2018
Task (short) Papers Due	21 May 2018
Acceptance Notification	28 May 2018
Camera-Ready Deadline	04 June 2018
Workshop Date	20 July 2018

3 Participation

1. Registration (12 March 2018). Prospective participants are to register through the NEWS 2018 website by requesting the datasets from 12 March onwards.

2. Train/Development Data (12 March 2018). Registered participants are to obtain train and development data from the shared task registration form and/or the designated copyright owners of databases. All registered participants are required to participate in the evaluation of at least one language pair, submit the results, prepare a short paper and attend the workshop at ACL 2018.

3. Test Data (07 May 2018). The test data would be released on 07 May 2018, and the participants have a maximum of 7 days to submit their results to the competition site. NEWS 2018 shared task will be run on CodaLab. Participants need to create a codalab account and register into the NEWS 2018 competition in order to be able to submit their system results. Only "standard" runs will be

processed this year. According to this, participants are required to use only the training and development data provided within the shared task to train their systems.

Participants can submit several runs for each individual language pair at the competition site. However, the total number of submissions per language pair will be limited to a maximum of 3 submissions per day, with a total maximum of 15 submissions during the whole period of the competition. From all submissions done to each individual language pair, each participant must select one to be posted on the leaderboard. Results on the leaderboard (by the last day of the shared task on 14 May 2018) will constitute the final official results of the shared task.

Each submission must be saved in a file named "results.xml" and submitted into the system in a ".zip" compressed file format. Each "results.xml" file can contain up to 10 output candidates in a ranked list for each corresponding input entry in the test file (refer to Appendix B for more details on file formating and naming conventions).

Those participants interested in submitting "non-standard" runs, i.e. transliteration results from systems that use additional data during the training phase, still will be able to do so. However such runs will be evaluated and reported separately (please contact the organizers).

4. Results (14 May 2018). Leaderboard results, as on 14 May 2018, will be considered the official evaluation results of the NEWS 2018 shared task. These results will be published on the workshop website and proceedings.

   Note that only the scores (evaluation metrics) of the participating systems on each language pair will be published, and no explicit reference to the participating teams will be provided. Furthermore, all participants should agree on not to reveal identities of other participants in any of their publications unless permission from the other respective participants is granted. By default, all participants remain anonymous in published results. Participating teams are allowed to reveal only their own identity in their publications.

5. Shared Task Short Papers (21 May 2018). Each participant is required to submit a 4-page system paper (short paper) describing their system, the used approach, submissions and results. Peer reviews will be conducted 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 workshop. All participants are required to register and attend the workshop to present their work. All paper submission and reviews will be managed electronically through https://www.softconf.com/acl2018/NEWS/.

4 Language Pairs

The different evaluation tasks within the NEWS 2018 shared task focus on transliteration and/or back-transliteration of personal and place names from a source language into a target language as summarized in Table 1. This year, the shared task offers 17 evaluation tasks, including 8 transliteration tasks, 5 back-transliteration tasks and 4 hybrid tasks. NEWS 2018 will release training, development and testing data for each of the language pairs. Within the 17 evaluation tasks, NEWS 2018 includes the 14 tasks that were evaluated in the previous year editions. In such cases, the training and development datasets are augmented versions of the previous year ones. New test dataset will be used in NEWS 2018 evaluations.

In order to estimate the progress of machine transliteration over time, the test/reference sets of NEWS 2016 have not been included in the training and development data for NEWS 2018. These test sets will be used as progress test sets to conduct a comparative study between NEWS 2018 and NEWS 2016 overall results.

The names given in the training sets for Thai, Persian, Chinese, Vietnamese, Hebrew, Japanese and Korean are Western names and their respective transliterations. The training set in the English to Japanese Kanji (B-JnJk) task consists only of native Japanese names. The training set in the Arabic to English (T-ArEn) task consists only of native Arabic names. Finally, the training sets for the English to Indian languages (Hindi, Tamil, Kannada and Bangla) tasks consist of a mix of both Indian and Western names.

Name origin	Source script	Target script	Type of Task	Dataset Size				Task ID
				Train	Dev	Prog	Test	
Western	English	Thai	Transliteration	30781	1000	1236	1000	T-EnTh
Western	Thai	English	Back-transliteration	27273	1000	1236	1000	B-ThEn
Western	English	Persian	Transliteration	13386	1000	1042	1000	T-EnPe
Western	Persian	English	Back-transliteration	15677	1000	-	1000	B-PeEn
Western	English	Chinese	Transliteration	41318	1000	1008	1000	T-EnCh
Western	Chinese	English	Back-transliteration	32002	1000	1019	1000	B-ChEn
Western	English	Vietnamese	Transliteration	3256	500	-	500	T-EnVi
Mixed	English	Hindi	Mixed trans/back	12937	1000	1000	1000	M-EnHi
Mixed	English	Tamil	Mixed trans/back	10957	1000	1000	1000	M-EnTa
Mixed	English	Kannada	Mixed trans/back	10955	1000	1000	1000	M-EnKa
Mixed	English	Bangla	Mixed trans/back	13623	1000	1000	1000	M-EnBa
Western	English	Hebrew	Transliteration	10501	1000	1100	1000	T-EnHe
Western	Hebrew	English	Back-transliteration	9447	1000	-	1000	B-HeEn
Western	English	Japanese Katakana	Transliteration	28828	1000	1033	1000	T-EnJa
Japanese	English	Japanese Kanji	Back-transliteration	10514	1000	1095	1000	B-JnJk
Western	English	Korean Hangul	Transliteration	7387	1000	1050	1000	T-EnKo
Arabic	Arabic	English	Transliteration	31354	1000	1156	1000	T-ArEn

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

5 Standard Datasets

Training Data (Parallel)

Paired names between source and target languages; size 3K – 41K.

Training data is used for training a basic transliteration system.

Development Data (Parallel)

Paired names between source and target languages; size 1K.

Development data is in addition to the training data, which is used for fine-tuning the system parameters, in case of need. Participants are allowed to use it as part of the training data for their final submissions.

Testing Data

Source names only; size 1K.

This is a held-out set, which will be used for evaluating the quality of the transliterations.

Progress Testing Data

Source names only; size 1K (approx).

This is the NEWS 2016 test set, it is held-out for progress comparative analysis.

Participants will need to obtain licenses from the respective copyright owners of the different datasets 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 2018 will provide the contact details for each dataset group.

The data would be provided in Unicode UTF-8 encoding, in XML format. The results are expected to be submitted in UTF-8 encoding also in XML format. The required XML format details are available in the Appendix A.

Note that name pairs are distributed as-is, as provided by the respective creators. While the datasets are mostly manually checked, there may be still inconsistencies (that is, non-standard usage, region-specific usage, errors, etc.) or incompleteness (that is, not all right variations may be covered). The participants are allowed to use any method of their preference to further clean up the data provided:

- For any participant conducting a manual clean up, 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!
- If automatic clean up were used, such clean up will be considered part of the system implementation, and hence it is not required to be shared with all participants.

All participants are required to use only the dataset (parallel names) provided by the shared task organizers for training their systems. This "standard" submission procedure will ensure a fair evaluation in term of score comparison across the different systems. Those participants wanting to additionally evaluate "non-standard" runs need to contact the organizers

6 Evaluation Metrics

As in previous editions of the shared task, the quality of the submitted results will be evaluated by using the following 4 metrics. Each individual name result might include up to 10 output candidates in a ranked list.

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 value, is taken for calculation. If the best matching reference is given by

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

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

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

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

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

- The length is computed in distinct Unicode characters.
- No distinction is made among 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

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

7 Paper Format

Paper submissions to NEWS 2018 should follow the ACL 2018 paper submission policy, including paper format, blind review policy and title and author conventions. Full papers (research papers) must be in two-column format without exceeding eight (8) pages of content plus two (2) extra pages for references and short papers (research and shared task papers) must also be in two-column format without exceeding four (4) pages content plus two (2) extra pages for references. Submission must conform to the official ACL 2018 style guidelines. For details, please refer to the ACL 2018 website: http://acl2018.org/call-for-papers/.

8 Contact Us

If you have any questions about the share task and the datasets, please contact any of the workshop organizers. Contact information is available at the NEWS 2018 website http://workshop.colips.org/news2018/contact.html

References

[CJKI2010] CJK Institute. 2010. http://www.cjk.org/.

[Li et al.2004] 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.

[MSRI2010] MSRI. 2010. Microsoft Research India. http://research.microsoft.com/india.

[AILab2018] Artificial Intelligence Laboratory (AILab) 2018. Ho Chi Minh City University of Science (VNU-HCMUS). https://www.ailab.hcmus.edu.vn/

[Cao et al.2010] Nam X. Cao, Nhut M. Pham, Quan H. Vu. 2010. Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model. In *Proc. Symposium on Information and Comunication Technology*, pages 59–63, ACM.

[Ngo et al.2015] Hoang Gia Ngo, Nancy F. Chen, Nguyen Binh Minh, Bin Ma, Haizhou Li. 2015. Phonology-Augmented Statistical Transliteration for Low-Resource Languages. Interspeech, 2015.

A Appendix: Data Formats

- File Naming Conventions:
  `NEWS18_Z-XXYY_trn.xml`
  `NEWS18_Z-XXYY_dev.xml`
  - `Z`: Type of task (T: transliteration, B: back-transliteration, M: mixed)
  - `XX`: Source Language
  - `YY`: Target Language
- File formats:
  All data will be made available in XML formats as illustrated in 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 Appendix: Submission of Results

- File Naming Conventions:
  Each submission must be saved in a file named "results.xml" and submitted into the NEWS 2018 CodaLab competition in a ".zip" compressed file. Each "results.xml" file can contain up to 10 output candidates in a ranked list for each corresponding input entry in the test file.
- File formats:
  All data will be provided in XML formats as illustrated in 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
    CorpusFormat = "UTF-8"
    CorpusID = "[task_id]"
    CorpusSize = "[total_number_of_names_in_file]"
    CorpusType = "[Training|Development]"
    NameSource = "[name_origin]"
    SourceLang = "[source_language]"
    TargetLang = "[target_language]">

    <Name ID="1">
        <SourceName>[source_name_1]</SourceName>
        <TargetName ID="1">[target_name_1_1]</TargetName>
        <TargetName ID="2">[target_name_1_2]</TargetName>
        ...
        <TargetName ID="n">[target_name_1_n]</TargetName>
    </Name>

    <Name ID="2">
        <SourceName>[source_name_2]</SourceName>
        <TargetName ID="1">[target_name_2_1]</TargetName>
        <TargetName ID="2">[target_name_2_2]</TargetName>
        ...
        <TargetName ID="k">[target_name_2_k]</TargetName>
    </Name>

    ...
    <!-- rest of the names to follow -->
    ...

</TransliterationCorpus>
```

Figure 1: Example of training and development data format.

```
<?xml version="1.0" encoding="UTF-8"?>

<TransliterationTaskResults
    SourceLang = "[source_language]"
    TargetLang = "[target_language]"
    GroupID = "[your_institution_name]"
    RunID = "[your_submission_number]"
    RunType = "Standard"
    Comments = "[your_comments_here]"
    TaskID = "[task_id]">

    <Name ID="1">
        <SourceName>[test_name_1]</SourceName>
        <TargetName ID="1">[your_system_result_1_1]</TargetName>
        <TargetName ID="2">[your_system_result_1_2]</TargetName>
        ...
        <TargetName ID="10">[your_system_result_1_10]</TargetName>
    </Name>

    <Name ID="2">
        <SourceName>[test_name_2]</SourceName>
        <TargetName ID="1">[your_system_result_2_1]</TargetName>
        <TargetName ID="2">[your_system_result_2_2]</TargetName>
        ...
        <TargetName ID="10">[your_system_result_2_10]</TargetName>
    </Name>

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

</TransliterationTaskResults>
```

Figure 2: Example of submission result format.