Report of NEWS 2016 Machine Transliteration Shared Task

Xiangyu Duan[2], Rafael E. Banchs[1], Min Zhang[2], Haizhou Li[1], A. Kumaran[3]

[1] Institute for Infocomm Research, A*STAR, Singapore 138632
{rembanchs,hli}@i2r.a-star.edu.sg

[2] Soochow University, China 215006
{xiangyuduan,minzhang}@suda.edu.cn

[3]Multilingual Systems Research, Microsoft Research India
a.kumaran@microsoft.com

Abstract

This report presents the results from the Machine Transliteration Shared Task conducted as part of The Sixth Named Entities Workshop (NEWS 2016) held at ACL 2016in Berlin, Germany. Similar to previous editions of NEWS Workshop, the Shared Task featured machine transliteration of proper names over 14 different language pairs, including 12 different languages and two different Japanese scripts.A total of 5 teams participated in the evaluation, submitting 255 standard and 19 non-standard runs, involving a diverse variety of transliteration methodologies. Four performance metrics were used to report the evaluation results. Once again, the NEWS shared task on machine transliteration has successfully achieved its objectives by providing a common ground for the research community to conduct comparative evaluations of state-of-the-art technologies that will benefit the future research and development in this area.

1 Introduction

Names play an important role in the performance of most Natural Language Processing (NLP) and Information Retrieval (IR) applications. They are also critical in cross-lingual applications such as Machine Translation (MT) and Cross-language Information Retrieval (CLIR), as it has been shown that system performance correlates positively with the quality of name conversion across languages (Demner-Fushman and Oard 2002, Mandl and Womser-Hacker 2005,Hermjakobet al. 2008, Udupa et al. 2009). Bilingual dictionaries constitute the traditional source of information for name conversion across languages, however they offer very limited support due to the fact that, in most languages, names are continuously emerging and evolving.

All of the above points to the critical need for robust Machine Transliteration methods and systems. During the last decade, significant efforts has been conducted by the research community to address the problem of machine transliteration (Knight and Graehl 1998, Meng et al. 2001, Li et al. 2004, Zelenko and Aone 2006, Sproat et al. 2006, Sherif and Kondrak 2007, Hermjakob et al. 2008, Al-Onaizan and Knight 2002,Goldwasser and Roth 2008, Goldberg and Elhadad 2008,Klementiev and Roth 2006, Oh and Choi 2002, Virga and Khudanpur 2003, Wan and Verspoor 1998, Kang and Choi 2000, Gao et al. 2004,Li et al. 2009a, Li et al. 2009b). These previous works fall into three main categories: grapheme-based, phoneme-based and hybrid methods. Graphemebased methods (Li et al. 2004) treat transliteration as a direct orthographic mapping and only uses orthography-related features while phoneme-based methods (Knight and Graehl 1998) make use of phonetic correspondences to generate the transliteration. The hybrid approach refers to the combination of several different models or knowledge sources to support the transliteration generation process.

The first machine transliteration shared task (Li et al. 2009b, Li et al. 2009a) was organized and conducted aspart of NEWS 2009 at ACL-IJCNLP 2009. It was the first time that common benchmarking data in diverse language pairs was provided for evaluating state-of-the-art machine transliteration. While the focus of the 2009 shared task was on establishing the quality metrics and on setting up a baselinefor transliteration quality based on those metrics, the 2010 shared task (Li et al. 2010a, Li et al. 2010b) focused on expanding the scope of the transliteration generation task to about a dozen languages and on exploring the quality of the task depending on the direction of transliteration. In NEWS 2011 (Zhang et al. 2011a, Zhang et al. 2011b),

the focus was on significantly increasing the hand-crafted parallel corpora of named entities to include 14 different language pairs from 11 language families, and on making them available as the common dataset for the shared task. The NEWS 2016Shared Task on Transliteration has been a continued effort for evaluating machine transliteration performance over such a common dataset following the NEWS 2015 (Banchs, et al., 2015), NEWS 2012and 2011 shared tasks.

In thispaper, we presentin full detail the results of the NEWS 2016Machine Transliteration Shared Task. The rest of the paper is structured as follows. Section 2 provides as short review of the main characteristics of the machine transliteration task and the corpora used for it. Section 3 reviews the four metrics used for the evaluations. Section 4 reports specific details about participation in the 2016 edition of the shared task, and section 5 presents and discusses the evaluation results. Finally, section 6 presents our main conclusions and future plans.

2 Shared Task on Transliteration

Transliteration, sometimes also called Romanization, especially if Latin Scripts are used for target strings (Halpern 2007), deals with the conversion of names between two languages and/or script systems. Within the context of the Transliteration Shared Task, we are aiming not only at addressing the name conversion process but also its practical utility for downstream applications, such as MT and CLIR.

In this sense, we adopt the same definition of transliteration as proposed during the NEWS 2009 workshop (Li et al. 2009a). According to it, transliteration is understood as the "conversion of a given name in the source language (a text string in the source writing system or orthography) to a name in the target language (another text string in the target writing system or orthography" conditioned to the following specific requirements regarding the name representation in the target language:

- it is phonetically equivalent to the source name,
- it conforms to the phonology of the target language, and
- it matches the user intuition on its equivalence with respect to the source language name.

Following NEWS 2011, NEWS 2012 and NEWS2015, the three back-transliteration tasks are maintained. Back-transliteration attempts to restore transliterated names back into their original source language. For instance, the tasks for converting western names written in Chinese and Thai back into their original English spellings are considered. Similarly, a task for back-transliterating Romanized Japanese names into their original Kanji strings is considered too.

2.1 Shared Task Description

Following the tradition of NEWS workshop series, the shared task in NEWS 2016 consists of developing machine transliteration systems in one or more of the specified language pairs.Each language pair of the shared task consists of asource and a target language, implicitly specifyingthe transliteration direction. Training and developmentdata in each of the language pairs was made available to all registered participants for developing their transliteration systems.

At the evaluation time, a standard hand-crafted test set consisting of between 500 and 3,000source names (approximately 5-10% of the trainingdata size) was released, on which theparticipants were required to produce a ranked listof transliteration candidates in the target languagefor each source name. The system output istested against a reference set (which may includemultiple correct transliterations for some sourcenames), and the performance of a system is capturedin multiple metrics (defined in Section 3),each designed to capture a specific performancedimension.

For every language pair, each participant was requiredto submit at least one run (designated as a"standard" run) that uses only the data provided bythe NEWS workshop organizers in that languagepair; i.e. no other data or linguistic resources are allowed for standard runs. Thisensures parity between systems andenables meaningful comparison of performanceof various algorithmic approaches in a given languagepair. Participants were allowed to submitone or more standard runs for each task they participated in. If more thanone standard runs were submitted, it was required toname one of them as a "primary" run, which was the one used to compare results across different systems.

In addition, more than one "non-standard" runs could besubmitted for every language pair using either databeyond theone provided by the shared task organizers,any other available linguistic resources in a specific language pair, orboth. This essentially enabled participants to demonstrate the limits of performance of theirsystems in a given language pair.

2.2 Shared Task Corpora

Two specific constraints were considered when selectinglanguages for the shared task: language diversityand data availability. To make the shared taskinteresting and to attract wider participation, it isimportant to ensure a reasonable variety amongthe languages in terms of linguistic diversity, orthographyand geography. Clearly, the ability ofprocuring and distributing a reasonably large (approximately10K paired names for training andtesting together) hand-crafted corpora consistingprimarily of paired names is critical for this process. Following NEWS 2015, the 14 tasks shown in Tables 1.a-ewere used (Li et al.2004, Kumaran and Kellner 2007, MSRI 2009,CJKI 2010).

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

For all of the tasks chosen, we have been able to procure paired-name data between thesource and the target scripts and were able tomake them available to the participants. Forsome language pairs, such as the case of English-Chinese andEnglish-Thai, there are both transliteration and-back-transliteration tasks. Most of the tasks are justone-way transliteration, although Indian data setscontainsa mixture of names from both Indian andWestern origins.

3 Evaluation Metrics and Rationale

The participants have been asked to submit standard and, optionally, non-standard runs.One of the standard runs must be named as the primary-submission, which was the one used for the performance summary.Each run must contain a ranked list of up toten candidate transliterations for each source name.The submitted results are compared to the groundtruth (reference transliterations) using four evaluationmetrics capturing different aspects of transliterationperformance. The four considered evaluation metrics are:

- Word Accuracy in Top-1 (ACC),
- Fuzziness in Top-1 (Mean F-score),
- Mean Reciprocal Rank (MRR), and
- Mean Average Precision (MAP_{ref}).

Task ID:EnCh			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Chinese	37K	2.8K	1.0K
Task ID: ChEn			data size		
Origin	Source	Target	Train	Dev	Test
Western	Chinese	English	28K	2.7K	1.0K

Table 1.a: Datasets provided by Institute for Infocomm Research, Singapore.

Task ID: EnKo			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Korean	7.0K	1.0K	1.0K
Task ID: EnJa			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Katakana	26K	2.0K	1.0K
Task ID: JnJk			data size		
Origin	Source	Target	Train	Dev	Test
Japanese	English	Kanji	10K	2.0K	1.0K
Task ID: ArEn			data size		
Origin	Source	Target	Train	Dev	Test
Arabic	Arabic	English	27K	2.5K	1.0K

Table 1.b: Datasets provided by the CJK Institute, Japan.

Task ID: EnHi			data size		
Origin	Source	Target	Train	Dev	Test
Mixed	English	Hindi	12K	1.0K	1.0K
Task ID: EnTa			data size		
Origin	Source	Target	Train	Dev	Test
Mixed	English	Tamil	10K	1.0K	1.0K
Task ID: EnKa			data size		
Origin	Source	Target	Train	Dev	Test
Mixed	English	Kannada	10K	1.0K	1.0K
Task ID: EnBa			data size		
Origin	Source	Target	Train	Dev	Test
Mixed	English	Bangla	13K	1.0K	1.0K
Task ID: EnHe			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Hebrew	9.5K	1.0K	1.0K

Table 1.c: Datasets provided by Microsoft Research India.

Task ID: EnTh			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Thai	27K	2.0K	1.0K
Task ID: ThEn			data size		
Origin	Source	Target	Train	Dev	Test
Western	Thai	English	25K	2.0K	1.0K

Table 1.d: Datasets provided by National Electronics and Computer Technology Center.

Task ID: EnPe			data size		
Origin	Source	Target	Train	Dev	Test
Western	English	Persian	10K	2.0K	1.0K

Table 1.e: Dataset provided bySarvnazKarimi / RMIT.

In the next subsections, we present a brief description of the four considered evaluation metrics. The following notation is further assumed:

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

3.1 Word Accuracy in Top-1 (ACC)

Also known as Word Error Rate, it measures correctnessof the first transliteration candidate in thecandidate list produced by a transliteration system.$ACC = 1$ means that all top candidates are correcttransliterations; i.e. they match one of the references,and $ACC = 0$ means that none of the topcandidates are correct.

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

3.2 Fuzziness in Top-1 (Mean F-score)

The Mean F-score measures how different, on average,the top transliteration candidate is from itsclosest reference. F-score for each source wordis a function of Precision and Recall and equals 1when the top candidate matches one of the references, and 0 when there are no common charactersbetween the candidate and any of the references.

Precision and Recall are calculated based onthe length of the Longest Common Subsequence(LCS) between a candidate and a reference:

$$LCS(c,r) = \frac{1}{2}\big(|c| + |r| - ED(c,r)\big) \quad \text{(Eq.2)}$$

whereED is the edit distance and $|x|$ is the lengthof x. For example, the longest common subsequencebetween "abcd" and "afcde" is "acd" andits length is 3. The best matching reference, i.e. the reference for which the edit distance hasthe minimum, is taken for calculation. If the best-matching reference is given by

$$r_{i,m} = \arg min_j \big(ED\big(c_{i,1}, r_{i,j}\big)\big) \quad \text{(Eq.3)}$$

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

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

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

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

The lengths are computed with respect to distinct Unicode characters, and no distinctions are made for different character types of a language (e.g. vowel vs. consonant vs. combining diereses, etc.).

3.3 Mean Reciprocal Rank (MRR)

Measures traditional MRR for any right answerproduced by the system, from among the candidates.1/MRR tells approximately the averagerank of the correct transliteration. MRR closer to 1implies that the correct answer is mostly producedclose to the top of the n-best lists.

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

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

3.4 Mean Average Precision (MAP$_{ref}$)

This metric measures tightly the precision in the n-best candidates for i-th source name, for which reference transliterations are available. If all of the referencesare produced, then the MAP is 1. If we denotethe number of correct candidates for the i-thsource word in k-best list as $num(i,k)$, then MAP$_{ref}$is given by:

$$MAP_{ref} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\big(\sum_{k=1}^{n_i} num(i,k)\big) \quad \text{(Eq.9)}$$

4 Participation in the Shared Task

A total of five teams from five different institutions participated in the NEWS 2016 Shared Task. More specifically, the participating teams were from National Institute of Information and Communications Technology (NICT), Qazvin Islamic Azad University (QIAU), University of Helsinki (UOH), Uppsala University (UPPS), Institute for Infocomm Research (I2R).

Teams were required to submit at least one standard run for every task they participated in, and for NEWS 2012/2015 test sets. They are set as the official NEWS 2016 evaluation set. In total, we received 31 standard and 2 non-standard runs for all test sets; i.e. 255 standard and 19 non-standard runs in total. Table 2 summarizes the number of standard runs, non-standard runs and teams participating per task.

Task	Std	Non	Teams Participating
EnCh	25	0	NICT,UPPS, QIAU, UOH
ChEn	27	2	NICT, UPPS, QIAU, UOH
EnKo	15	0	NICT
EnJa	15	0	NICT
JnJk	15	0	NICT
ArEn	0	0	
EnHi	19	0	NICT, QIAU, UOH
EnTa	17	0	NICT, QIAU
EnKa	17	0	NICT, QIAU
EnBa	20	0	NICT, QIAU, UOH
EnHe	24	2	NICT, QIAU, UOH
EnTh	22	0	NICT, QIAU, I2R
ThEn	16	0	NICT, QIAU
EnPe	23	15	NICT, QIAU, UOH
	255	19	

Table 2: Number of standard (Std) and non-standard (Non) runs submitted, and teams participating in each task.

As seen from the table, the most popular task continues to be the transliteration from English to Chinese and Chinese to English (Zhang et al. 2012), followed by English to Hindi etc. Non-standard runs were only submitted for 3 of the 14 tasks.

4.1 Shared Task on CodaLab

Different from previous years, in NEWS 2016 the Shared Task evaluation was run online by using the CodaLab platform (http://codalab.org/). CodaLab is a powerful online platform aiming at accelerating reproducible computational research. Two main functionalities are available at the CodaLab platform: worksheets, which allows for running reproducible experiments and creating executable papers; and competitions, which allows for participating and/or hosting competitions.

CodaLab's competitions allows for running competitions that involve either code submissions or data submissions. For the case of NEWS 2016 Shared Task on transliteration, two CodaLab competitions on the data submission modality were created: NEWS 2016 Standard submissions (https://competitions.codalab.org/ competitions/8991) and NEWS 2016 Non-standard submissions(https://competitions.codalab.org/ competitions/9021). In the standard submissions competition, participants were required to use only the training and development data provided by the Shared Task, while for the non-standard submissions competitions, in addition to the training and development data provided by the Shared Task, participants were welcomed to use external data, either parallel or monolingual. A total of 12 and 4 participants registered for the standard submissions and non-standard submissions competitions, respectively, but finally only five teams submitted results into the competitions.

Each competition was composed of 14 phases, each corresponding to one of the 14 transliteration tasks available in the Shared Task. All phases were run in parallel, meaning that each participant was able to submit results to any of the phases at any moment during the evaluation campaign, which ran from April 25th to May 3rd. During this period, participants were allowed to submit to each of the two competitions up to 3 results per day and per task, with an overall maximum of 15 submissions per task during the complete evaluation period. For each task they participated in, participants were allowed to post only one result in the corresponding leader-board. The leader-boards for both the standard submissions and non-standard submissions competitions are available at https://competitions.codalab.org/ competitions/8991#results and https:// competitions.codalab.org/competitions/9021#results, respectively.

4.2 Baseline System Results

Also different from previous years, in NEWS 2016 a baseline system was set up and baseline results were computed for all the 14 transliteration tasks available in the Shared Task. Baseline results were based on a simple MT implementation at the character level using MOSES. The baseline system was generously provided by UPC, Barcelona (Costa-jussa, 2016).

A summary of NEWS 2016 Shared Task results, including the MOSES-based baseline results, is available in the workshop's website at: http://workshop.colips.org/news2016/results.html.

5 Task Results and Analysis

Figure 1 summarizes the results of the NEWS 2016 Shared Task. In the figure, only F-scores over the NEWS 2015 evaluation test set (referred to as NEWS15/16) for all primary standard submissions are depicted. A total of 31 primary standard submissions were received.

As seen from the figure, with the exception of the English to Japanese Katakana, only transliteration tasks involving Arabic, Persian and the four considered Indian languages are consistently scored above 80%. For the rest of the languages, with the exception of Japanese Katakana and Hebrew, scores are consistently in the range from 60% to 80%. Notice also that, regardless the availability of training data, the English to Chinese transliteration task seems to be the more demanding one for state-of-the-art systems with respect to the considered metric.

Another interesting observation that can be derived from the figure, when looking to the language pairs English-Chinese and English-Thai, is that systems tend to perform slightly better for the case of back-transliteration tasks.

A much more comprehensive presentation of results for the NEWS 2016 Shared Task is provided in the Appendix at the end of this paper. There, resulting scores are reported for all received submissions, including standard and non-standard submissions and the four considered evaluation metrics. All results are presented in 14 tables, each of which reports the scores for one transliteration task over one test set. In the tables, all primary standard runs are highlighted in bold-italic fonts.

Regarding the systems participating in this year evaluation, two highest performance systems of the five participants submitted their system description papers, which are from NICT and UPPS. The NICT's system (Finch et al. 2016) applied neural network Ensembles, each of which explores the agreement of target-bidirectional sequence-to-sequence neural network model. The ensembles show great improvements over their NEWS 2015 results, which utilized a rescoring reranking function to ensemble attention-based neural network and traditional machine translation models.

The UPPS's system (Shao et al. 2016) implemented a neural network trained on unsupervised sub-units alignments. They used a convolutional neural network to encode character-level transliteration information and a recurrent neural network as stacking. Their decoding performance demonstrates that their proposed neural network significantly outperforms the baseline which is a character-level system trained by Moses.

As seen from the previous system descriptions, neural networks become more and more predominant in the state-of-the-art machine transliteration. Significant improvements are achieved by neural network ensembles, while single neural

network also obtains better performance than traditional phrase-based machine translation systems. The simple ensemble method achieved the best performance across all 14 phases.
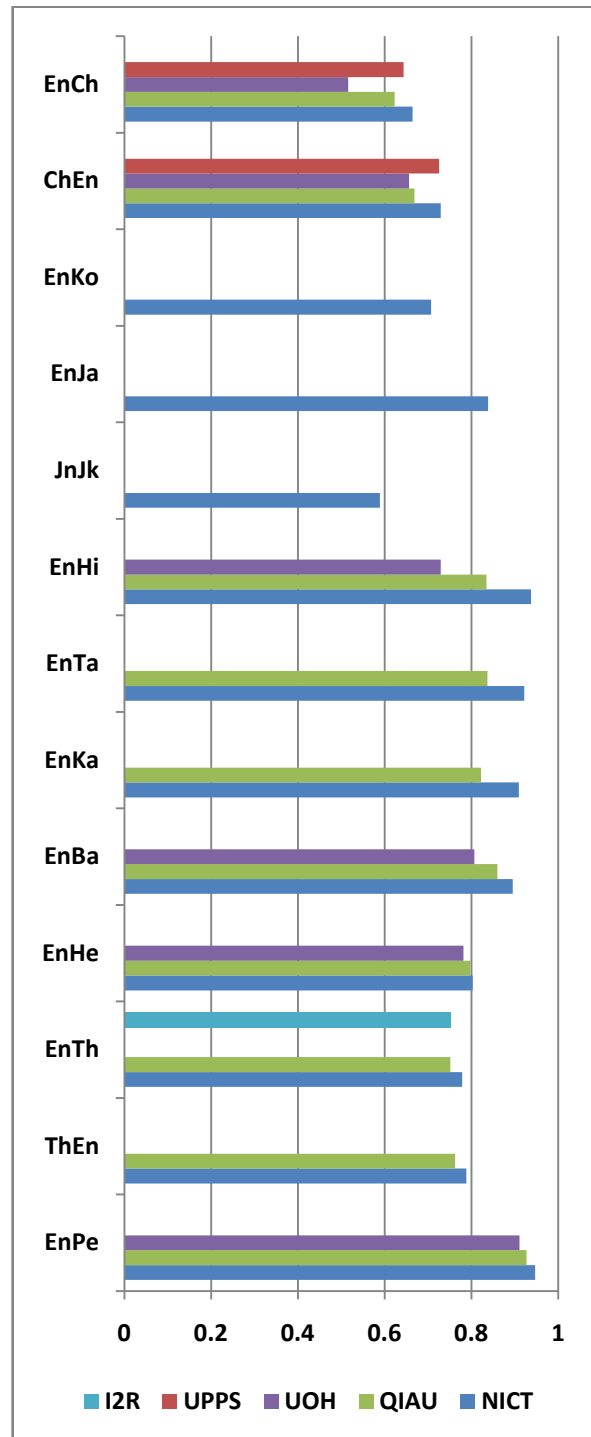

Figure 1: Mean F-scores (Top-1) on the evaluation test set (NEWS12/15) for all primary standard submissions and all transliteration tasks.

Finally, figure 2 compares, in terms of Mean F-scores, the best primary standard submissions in NEWS 2015 with the ones in NEWS 2016.
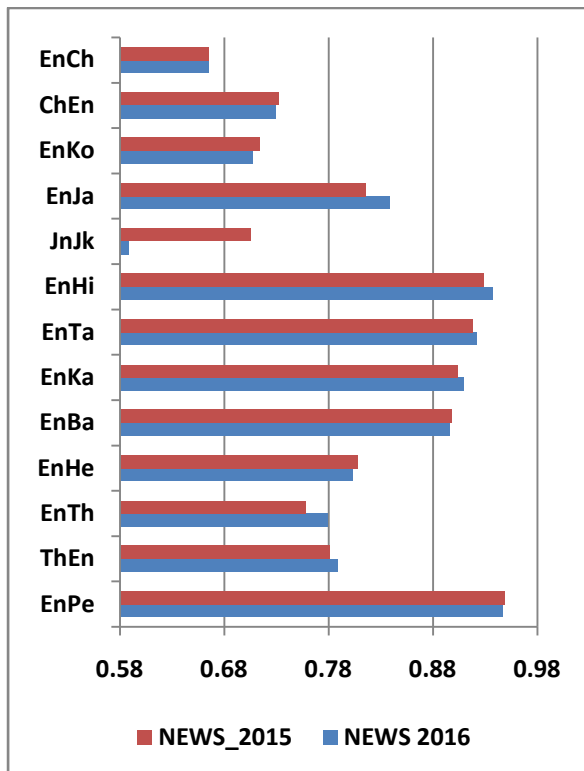
Figure 2: Mean F-scores (Top-1) on the evaluation test set (NEWS12/15) for the best primary standard submissions in 2012 and 2015.

As seen from the figure, in most of the considered transliteration tasks, some incremental improvements can be observed between the 2015 and 2016 shared tasks. The most significant improvements are in those tasks involving Japanese Katakana, Tamil, Kannada, and Thai.

Regarding the observed drops in performance, the most significant one is from JnJk. It is mainly due to that the specific participant NICT applied a totally different methodology compared to JnJk in 2015. As their system description paper points out, the drop is because the large vocabulary set on the target side that neural network hardly handles.

6 Conclusions

The Shared Task on Machine Transliteration in NEWS 2016 has shown, once again, that the research community has a continued interest in this area. This report summarizes the results of the NEWS 2016 Shared Task.

We are pleased to report a comprehensive set of machine transliteration approaches and their evaluation results over the evaluation test set, as well as two conditions: standard runs and non-standard runs.While the standard runs allow for conducting meaningful comparisons across different algorithms, the non-standard runs open up more opportunities for exploiting a varietyof additional linguistic resources.

Five teams from five different institutions participated in the shared task. In total, we received 31 standard and 2 non-standard runs for each test set; i.e. 255 standard and 19 non-standard runs in total. Most of the current state-of-the-art in machine transliteration is represented in the systems that have participated in the shared task.

Encouraged by the continued success of the NEWS workshopseries, we plan to continue this event in the future to further promoting machine transliteration research and development.

Acknowledgments

References

Y. Al-Onaizan, K. Knight. 2002. Machine transliteration of names in arabic text. In Proc. ACL-2002Workshop: Computational Apporaches to Semitic Languages, Philadelphia, PA, USA.

D. Bahdanau, K. Cho, Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cornell University Library, arXiv:1409.0473 [cs.CL]

R. Banchs; M. Zhang; X. Duan; H. Li; A. Kumaran. 2015. Report of NEWS 2015 Machine Transliteration Shared Task. Proceedings of the Fifth Named Entity Workshop, joint with 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China.

M. Bisani, H. Ney. 2008. Joint sequence models for grapheme-to-phoneme conversion. Speech Communication, 50(5):434–451.

CJKI. 2010. CJK Institute. http://www.cjk.org/.

M. Costa-juss`. 2016. Moses-based official baseline for NEWS 2016. Proceedings of the Sixth Named Entity Workshop, joint with 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

D. Demner-Fushman, D.W. Oard. 2002. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In Proc. 36-th Hawaii Int'l. Conf. System Sciences, volume 4, page 108.2.

A. Finch, P. Dixon, E. Sumita. 2012. Rescoring a phrase-based machine transliteration system with recurrent neural network language models. In Proceedings of the 4th Named Entity Workshop (NEWS) 2012, pages 47–51, Jeju, Korea, July.

A. Finch, L. Liu, X. Wang, E. Sumita. 2015.Neural Network Transduction Models in Transliteration Generation. In Proceedings of the 2015 Named Entities Workshop: Shared Task on Transliteration (NEWS 2015), Beijing, China.

A. Finch, L. Liu, X. Wang, E. Sumita. 2016. Target-Bidirectional Neural Models for Machine Transliteration. In Proceedings of the 2016 Named Entities Workshop: Shared Task on Transliteration (NEWS 2016), Berlin, German.

W. Gao, K.F. Wong, W. Lam. 2004. Phoneme-based transliteration of foreign names for OOV problem. In Proc. IJCNLP, pages 374–381, Sanya, Hainan, China.

Y. Goldberg, M. Elhadad. 2008. Identification of transliterated foreign words in Hebrew script. In Proc. CICLing, volume LNCS 4919, pages 466–477.

D. Goldwasser, D. Roth. 2008. Transliteration as constrained optimization. In Proc. EMNLP, pages 353–362.

J. Halpern. 2007. The challenges and pitfalls of Arabic romanization and arabization. In Proc. Workshop on Comp. Approaches to Arabic Scriptbased Lang.

U. Hermjakob, K. Knight, H. Daum. 2008. Name translation in statistical machine translation: Learning when to transliterate. In Proc. ACL, Columbus, OH, USA, June.

S. Jiampojamarn, G. Kondrak, T. Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; pages 372–379, Rochester, New York, April.

S. Jiampojamarn, A. Bhargava, Q. Dou, K. Dwyer, G. Kondrak. 2009. DirecTL: a language independent approach to transliteration. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pages 28–31, Suntec, Singapore.

S. Jiampojamarn, C. Cherry, G. Kondrak. 2010. Integrating joint n-gram features into a discriminative training framework. In Proceedings of NAACL-2010, Los Angeles, CA, June.Association for Computational Linguistics.

B.J. Kang, K.S. Choi. 2000. English-Korean automatic transliteration/ backtransliteration system and character alignment. In Proc. ACL, pages 17–18, Hong Kong.

A. Klementiev, D. Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In Proc. 21st Int'l Conf Computational Linguistics and 44th Annual Meeting of ACL, pages 817–824, Sydney, Australia, July.

K. Knight, J. Graehl. 1998. Machine transliteration. Computational Linguistics, 24(4).

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.

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

A. Kunchukuttan, P. Bhattacharyya. 2015. Data representation methods and use of mined corpora for Indian language transliteration. In Proceedings of the 2015 Named Entities Workshop: Shared Task on Transliteration (NEWS 2015), Beijing, China.

H. Li, M. Zhang, J. Su. 2004. A joint source-channel model for machine transliteration. In Proc. 42nd ACL Annual Meeting, pages 159–166, Barcelona, Spain.

H. Li, A. Kumaran, V. Pervouchine, M. Zhang. 2009a. Report of NEWS 2009 machine transliteration shared task. In Proc. Named Entities Workshop at ACL 2009.

H. Li, A. Kumaran, M. Zhang, V. Pervouchine. 2009b. ACL-IJCNLP 2009 Named Entities Workshop - Shared Task on Transliteration. In Proc. Named Entities Workshop at ACL 2009.

H. Li, A. Kumaran, M. Zhang, V. Pervouchine. 2010a. Report of news 2010 transliteration generation shared task. In Proc. Named Entities Workshop at ACL 2010.

H. Li, A. Kumaran, M. Zhang, V. Pervouchine. 2010b. Whitepaper of news 2010 shared task on transliteration generation. In Proc. Named Entities Workshop at ACL 2010.

T. Mandl, C. Womser-Hacker. 2005. The effect of named entities on effectiveness in cross-language information retrieval evaluation. In Proc. ACM Symp.Applied Comp., pages 1059–1064.

H.M. Meng, W.K. Lo, B. Chen, K. Tang. 2001. Generate phonetic cognates to handle name entities in English-Chinese cross-language spoken document retrieval. In Proc. ASRU.

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

G. Nicolai, B. Hauer, M. Salameh, A. St Arnaud, Y. Xu, L. Yao, G. Kondrak. 2015. Multiple System Combination for Transliteration. In Proceedings of the 2015 Named Entities Workshop: Shared Task on Transliteration (NEWS 2015), Beijing, China.

J.H. Oh, K.S. Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In Proc. COLING 2002, Taipei, Taiwan.

Y. Shao, J. Nivre. 2016. Applying Neural Networks to English-Chinese Named Entity Transliteration. In Proceedings of the 2016 Named Entities Workshop: Shared Task on Transliteration (NEWS 2016), Berlin, Germany.

T. Sherif, G. Kondrak. 2007. Substringbased transliteration. In Proc. 45th Annual Meeting of the ACL, pages 944–951, Prague, Czech Republic, June.

R. Sproat, T. Tao, C.X. Zhai. 2006. Named entity transliteration with comparable corpora. In Proc. 21st Int'l Conf Computational Linguistics and 44th Annual Meeting of ACL, pages 73–80, Sydney, Australia.

R. Udupa, K. Saravanan, A. Bakalov, A. Bhole. 2009. "They are out there, if you know where to look": Mining transliterations of OOV query terms for cross-language information retrieval. In LNCS: Advances in Information Retrieval, volume 5478, pages 437–448. Springer Berlin / Heidelberg.

P. Virga, S. Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In Proc. ACL MLNER, Sapporo, Japan.

S. Wan, C.M. Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In Proc. COLING, pages 1352–1356.

D. Wang, X. Yang, J. Xu, Y. Chen, N. Wang, B. Liu, J. Yang, Y. Zhang. 2015a. A Hybrid Transliteration Model for Chinese/English Named Entities — BJTU-NLP Report for the 5th Named Entities Workshop. In Proceedings of the 2015 Named Entities Workshop: Shared Task on Transliteration (NEWS 2015), Beijing, China.

Y.C. Wang, C.K. Wu, R.T.H. Tsai. 2015b. NCU IISR English-Korean and English-Chinese Named Entity Transliteration Using Different Grapheme Segmentation Approaches. In Proceedings of the 2015 Named Entities Workshop: Shared Task on Transliteration (NEWS 2015), Beijing, China.

D. Zelenko, C. Aone. 2006. Discriminative methods for transliteration. In Proc. EMNLP, pages 612–617, Sydney, Australia, July.

M. Zhang, A. Kumaran, H. Li. 2011a. Whitepaper of news 2011 shared task on machine transliteration. In Proc. Named Entities Workshop at IJCNLP 2011.

M. Zhang, H. Li, A. Kumaran, M. Liu. 2011b. Report of news 2011 machine transliteration shared task. In Proc. Named Entities Workshop at IJCNLP 2011.

M. Zhang, H. Li, A. Kumaran, M. Liu. 2012. Report of NEWS 2012 Machine Transliteration Shared Task. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 10–20, Jeju, Republic of Korea.

Appendix: Evaluation Results

Team	Accuracy	F-score	MRR	MAP$_{ref}$
UPPS	0.3353 (1)	0.6759 (1)	0.3963 (4)	0.3233 (1)
NICT	0.3165 (2)	0.6643 (2)	0.4130 (1)	0.3086 (2)
NICT	0.3145 (3)	0.6629 (3)	0.4100 (2)	0.3056 (3)
NICT	0.3085 (4)	0.6604 (5)	0.4037 (3)	0.2995 (4)
NICT	0.3056 (5)	0.6467 (14)	0.3938 (5)	0.2946 (6)
NICT	0.3006 (6)	0.6606 (4)	0.3920 (6)	0.2949 (5)
NICT	0.2996 (7)	0.6489 (13)	0.3871 (7)	0.2924 (7)
NICT	0.2986 (8)	0.6533 (10)	0.3739 (11)	0.2901 (8)
NICT	0.2966 (9)	0.6540 (9)	0.3814 (8)	0.2873 (10)
NICT	0.2966 (9)	0.6567 (6)	0.3799 (9)	0.2858 (11)
NICT	0.2966 (9)	0.6564 (7)	0.3787 (10)	0.2887 (9)
NICT	0.2877 (10)	0.6509 (11)	0.3698 (14)	0.2798 (13)
NICT	0.2867 (11)	0.6546 (8)	0.3724 (12)	0.2761 (14)
NICT	0.2837 (12)	0.6491 (12)	0.3711 (13)	0.2750 (15)
UPPS	0.2808 (13)	0.6434 (15)	0.3007 (18)	0.2812 (12)
NICT	0.2768 (14)	0.6421 (16)	0.3591 (15)	0.2680 (16)
NICT	0.2728 (15)	0.6373 (17)	0.3583 (16)	0.2653 (17)
QIAU	0.2659 (16)	0.6227 (18)	0.3185 (17)	0.2549 (18)
UOH	0.1062 (17)	0.5160 (19)	0.1062 (19)	0.1014 (19)
UOH	0.0992 (18)	0.5021 (20)	0.0992 (20)	0.0945 (20)
UOH	0.0010 (19)	0.1989 (21)	0.0010 (21)	0.0010 (21)
UOH	0.0010 (19)	0.1989 (21)	0.0010 (21)	0.0010 (21)
QIAU	0.0000 (20)	0.0871 (22)	0.0000 (22)	0.0000 (22)

Table 1: Results for the English to Chinese transliteration task (EnCh) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.2139 (1)	0.7292 (4)	0.3026 (2)	0.2119 (1)
NICT	0.2110 (2)	0.7309 (2)	0.3016 (3)	0.2091 (2)
NICT	0.2100 (3)	0.7308 (3)	0.3038 (1)	0.2076 (3)
NICT	0.2012 (4)	0.7212 (6)	0.2774 (6)	0.1978 (4)
NICT	0.2002 (5)	0.7182 (7)	0.2798 (5)	0.1978 (5)
UPPS	0.1992 (6)	0.7524 (1)	0.2810 (4)	0.1947 (6)
NICT	0.1904 (7)	0.7067 (14)	0.2679 (9)	0.1866 (7)
NICT	0.1884 (8)	0.7150 (10)	0.2718 (7)	0.1831 (10)
NICT	0.1884 (8)	0.7165 (8)	0.2693 (8)	0.1854 (8)
NICT	0.1865 (9)	0.7164 (9)	0.2678 (10)	0.1851 (9)
NICT	0.1835 (10)	0.7103 (13)	0.2658 (11)	0.1813 (11)
NICT	0.1825 (11)	0.7120 (12)	0.2606 (12)	0.1800 (12)
NICT	0.1766 (12)	0.7122 (11)	0.2550 (14)	0.1735 (13)
NICT	0.1747 (13)	0.7065 (15)	0.2560 (13)	0.1717 (14)
NICT	0.1698 (14)	0.6987 (17)	0.2481 (15)	0.1679 (15)
UPPS	0.1619 (15)	0.7253 (5)	0.1816 (17)	0.1620 (16)
UPPS	0.1619 (15)	0.7253 (5)	0.1816 (17)	0.1620 (16)
NICT	0.1600 (16)	0.6989 (16)	0.2450 (16)	0.1578 (17)

UOH	0.1099 (17)	0.6687 (18)	0.1099 (19)	0.1066 (18)
UOH	0.0854 (18)	0.6500 (20)	0.0854 (20)	0.0825 (20)
QIAU	0.0834 (19)	0.6564 (19)	0.1425 (18)	0.0830 (19)
QIAU	0.0834 (19)	0.6564 (19)	0.1425 (18)	0.0830 (19)
UOH	0.0000 (20)	0.4770 (21)	0.0000 (21)	0.0000 (21)
UOH	0.0000 (20)	0.4770 (21)	0.0000 (21)	0.0000 (21)

Table 2: Results for the Chinese to English transliteration task (ChEn) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP_{ref}
NICT	0.1869 (1)	0.7784 (1)	0.2789 (1)	0.1869 (1)
NICT	0.1837 (2)	0.7759 (2)	0.2782 (2)	0.1837 (2)
NICT	0.1788 (3)	0.7735 (3)	0.2740 (3)	0.1788 (3)
NICT	0.1756 (4)	0.7626 (6)	0.2575 (4)	0.1756 (4)
NICT	0.1707 (5)	0.7675 (4)	0.2501 (6)	0.1707 (5)
NICT	0.1691 (6)	0.7557 (8)	0.2384 (8)	0.1691 (6)
NICT	0.1650 (7)	0.7635 (5)	0.2501 (5)	0.1650 (7)
NICT	0.1586 (8)	0.7567 (7)	0.2439 (7)	0.1586 (8)
NICT	0.1570 (9)	0.7550 (10)	0.2382 (10)	0.1570 (9)
NICT	0.1570 (9)	0.7550 (9)	0.2382 (9)	0.1570 (9)
NICT	0.1553 (10)	0.7550 (11)	0.2372 (11)	0.1553 (10)
I2R	0.1553 (10)	0.7537 (13)	0.1561 (17)	0.1553 (10)
NICT	0.1537 (11)	0.7516 (15)	0.2297 (14)	0.1537 (11)
NICT	0.1529 (12)	0.7542 (12)	0.2346 (12)	0.1529 (12)
NICT	0.1456 (13)	0.7497 (17)	0.2229 (15)	0.1456 (13)
QIAU	0.1456 (13)	0.7514 (16)	0.2181 (16)	0.1456 (13)
NICT	0.1448 (14)	0.7522 (14)	0.2314 (13)	0.1448 (14)
I2R	0.1440 (15)	0.7491 (18)	0.1448 (18)	0.1440 (15)
I2R	0.1173 (16)	0.7452 (19)	0.1173 (19)	0.1173 (16)
I2R	0.0631 (17)	0.6687 (20)	0.0631 (20)	0.0631 (17)
I2R	0.0437 (18)	0.6572 (21)	0.0437 (21)	0.0437 (18)

Table 3: Results for the English to Thai transliteration task (EnTh) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP_{ref}
NICT	0.1958 (1)	0.7881 (2)	0.2914 (1)	0.1958 (1)
NICT	0.1942 (2)	0.7897 (1)	0.2885 (2)	0.1942 (2)
NICT	0.1942 (2)	0.7871 (3)	0.2853 (3)	0.1942 (2)
NICT	0.1699 (3)	0.7739 (4)	0.2531 (4)	0.1699 (3)
NICT	0.1667 (4)	0.7692 (6)	0.2424 (5)	0.1667 (4)
NICT	0.1594 (5)	0.7681 (8)	0.2413 (6)	0.1594 (5)
NICT	0.1537 (6)	0.7659 (11)	0.2326 (10)	0.1537 (6)
NICT	0.1529 (7)	0.7657 (13)	0.2325 (11)	0.1529 (7)
NICT	0.1521 (8)	0.7681 (9)	0.2412 (7)	0.1521 (8)
NICT	0.1521 (8)	0.7691 (7)	0.2354 (8)	0.1521 (8)
NICT	0.1513 (9)	0.7630 (15)	0.2207 (15)	0.1513 (9)
NICT	0.1497 (10)	0.7652 (14)	0.2245 (13)	0.1497 (10)
NICT	0.1472 (11)	0.7673 (10)	0.2280 (12)	0.1472 (11)
NICT	0.1464 (12)	0.7657 (12)	0.2221 (14)	0.1464 (12)

NICT	0.1456 (13)	0.7715 (5)	0.2330 (9)	0.1456 (13)
QIAU	0.1286 (14)	0.7624 (16)	0.1966 (16)	0.1286 (14)

Table 4: Results for the Thai to English transliteration task (ThEn) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.6958 (1)	0.9466 (2)	0.7952 (1)	0.6799 (1)
NICT	0.6939 (2)	0.9481 (1)	0.7921 (2)	0.6787 (2)
NICT	0.6910 (3)	0.9455 (3)	0.7908 (3)	0.6747 (3)
NICT	0.6555 (4)	0.9388 (5)	0.7594 (6)	0.6388 (5)
NICT	0.6545 (5)	0.9391 (4)	0.7597 (5)	0.6402 (4)
NICT	0.6488 (6)	0.9373 (6)	0.7627 (4)	0.6326 (6)
NICT	0.6449 (7)	0.9363 (8)	0.7562 (8)	0.6288 (7)
NICT	0.6401 (8)	0.9363 (7)	0.7564 (7)	0.6245 (9)
NICT	0.6382 (9)	0.9356 (10)	0.7473 (10)	0.6199 (10)
NICT	0.6382 (9)	0.9357 (9)	0.7531 (9)	0.6273 (8)
NICT	0.6296 (10)	0.9351 (11)	0.7463 (11)	0.6153 (11)
NICT	0.6276 (11)	0.9328 (12)	0.7385 (12)	0.6137 (12)
NICT	0.5912 (12)	0.9299 (13)	0.7174 (13)	0.5763 (13)
QIAU	0.5816 (13)	0.9267 (14)	0.7116 (14)	0.5673 (15)
QIAU	0.5816 (13)	0.9267 (14)	0.7116 (14)	0.5673 (15)
NICT	0.5797 (14)	0.9235 (16)	0.7012 (16)	0.5710 (14)
QIAU	0.5758 (15)	0.9261 (15)	0.7090 (15)	0.5633 (16)
UOH	0.5077 (16)	0.9103 (18)	0.5077 (17)	0.4718 (17)
UOH	0.5029 (17)	0.9104 (17)	0.5029 (18)	0.4678 (18)
UOH	0.3369 (18)	0.8711 (19)	0.3369 (19)	0.3138 (19)
NICT	0.0000 (19)	0.0000 (20)	0.0000 (20)	0.0000 (20)

Table 5: Results for the English to Persian transliteration task (EnPe) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.7150 (1)	0.9371 (1)	0.7814 (2)	0.7091 (1)
NICT	0.7130 (2)	0.9339 (3)	0.7821 (1)	0.7086 (2)
NICT	0.7090 (3)	0.9344 (2)	0.7807 (3)	0.7044 (3)
NICT	0.6460 (4)	0.9174 (5)	0.7223 (7)	0.6425 (4)
NICT	0.6450 (5)	0.9164 (7)	0.7249 (5)	0.6382 (5)
NICT	0.6410 (6)	0.9162 (8)	0.7265 (4)	0.6356 (6)
NICT	0.6370 (7)	0.9160 (9)	0.7172 (9)	0.6306 (7)
NICT	0.6370 (7)	0.9191 (4)	0.7241 (6)	0.6305 (8)
NICT	0.6310 (8)	0.9164 (6)	0.7181 (8)	0.6264 (9)
NICT	0.6270 (9)	0.9103 (11)	0.7125 (10)	0.6229 (10)
NICT	0.6230 (10)	0.9138 (10)	0.7116 (11)	0.6190 (11)
NICT	0.6120 (11)	0.9090 (13)	0.7066 (12)	0.6085 (12)
NICT	0.6080 (12)	0.9098 (12)	0.6994 (14)	0.6037 (13)
NICT	0.6050 (13)	0.9057 (14)	0.6995 (13)	0.5992 (14)
NICT	0.5780 (14)	0.9055 (15)	0.6814 (15)	0.5724 (15)
QIAU	0.3480 (15)	0.8349 (16)	0.4745 (16)	0.3434 (16)
UOH	0.1090 (16)	0.7292 (17)	0.1090 (17)	0.1061 (17)
UOH	0.0300 (17)	0.6073 (18)	0.0300 (18)	0.0285 (18)

Team	Accuracy	F-score	MRR	MAP_ref
QIAU	0.0000 (18)	0.2553 (19)	0.0002 (19)	0.0000 (19)

Table 6: Results for the English to Hindi transliteration task (EnHi) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP_ref
NICT	0.6290 (1)	0.9216 (1)	0.7171 (1)	0.6280 (1)
NICT	0.6200 (2)	0.9204 (2)	0.7148 (2)	0.6190 (2)
NICT	0.6130 (3)	0.9186 (3)	0.7120 (3)	0.6120 (3)
NICT	0.5910 (4)	0.9101 (4)	0.6795 (4)	0.5902 (4)
NICT	0.5890 (5)	0.9091 (5)	0.6778 (5)	0.5882 (5)
NICT	0.5720 (6)	0.9074 (6)	0.6674 (7)	0.5712 (6)
NICT	0.5690 (7)	0.9048 (7)	0.6725 (6)	0.5685 (7)
NICT	0.5640 (8)	0.9029 (8)	0.6597 (8)	0.5642 (8)
NICT	0.5450 (9)	0.8999 (10)	0.6500 (9)	0.5445 (9)
NICT	0.5390 (10)	0.8986 (11)	0.6399 (11)	0.5388 (10)
NICT	0.5380 (11)	0.9011 (9)	0.6465 (10)	0.5370 (11)
NICT	0.5180 (12)	0.8949 (12)	0.6248 (12)	0.5182 (12)
NICT	0.4980 (13)	0.8883 (13)	0.6081 (14)	0.4972 (13)
NICT	0.4950 (14)	0.8880 (14)	0.6099 (13)	0.4940 (14)
NICT	0.4460 (15)	0.8824 (15)	0.5731 (15)	0.4452 (15)
QIAU	0.3240 (16)	0.8369 (16)	0.4461 (16)	0.3235 (16)
QIAU	0.0000 (17)	0.3623 (17)	0.0001 (17)	0.0000 (17)

Table 7: Results for the English to Tamil transliteration task (EnTa) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP_ref
NICT	0.5830 (1)	0.9089 (1)	0.6815 (1)	0.5819 (1)
NICT	0.5700 (2)	0.9076 (2)	0.6752 (2)	0.5689 (2)
NICT	0.5650 (3)	0.9053 (3)	0.6703 (3)	0.5639 (3)
NICT	0.5250 (4)	0.8894 (6)	0.6159 (6)	0.5246 (4)
NICT	0.5230 (5)	0.8902 (4)	0.6229 (4)	0.5219 (5)
NICT	0.5190 (6)	0.8900 (5)	0.6170 (5)	0.5185 (6)
NICT	0.5130 (7)	0.8832 (11)	0.6134 (8)	0.5128 (7)
NICT	0.5070 (8)	0.8861 (10)	0.6078 (10)	0.5061 (9)
NICT	0.5070 (8)	0.8892 (7)	0.6129 (9)	0.5059 (10)
NICT	0.5070 (8)	0.8867 (9)	0.6156 (7)	0.5066 (8)
NICT	0.5020 (9)	0.8870 (8)	0.6022 (11)	0.5011 (11)
NICT	0.4790 (10)	0.8801 (12)	0.5864 (12)	0.4792 (12)
NICT	0.4320 (11)	0.8695 (13)	0.5569 (13)	0.4314 (13)
NICT	0.4280 (12)	0.8635 (14)	0.5479 (14)	0.4280 (14)
NICT	0.3990 (13)	0.8509 (15)	0.5001 (15)	0.3988 (15)
QIAU	0.2860 (14)	0.8224 (16)	0.4019 (16)	0.2856 (16)
QIAU	0.0000 (15)	0.3288 (17)	0.0000 (17)	0.0000 (17)

Table 8: Results for the English to Kannada transliteration task (EnKa) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP_ref
NICT	0.4980 (1)	0.8955 (3)	0.6266 (1)	0.4978 (1)
NICT	0.4980 (1)	0.8975 (1)	0.6243 (2)	0.4965 (2)
NICT	0.4890 (2)	0.8963 (2)	0.6167 (3)	0.4873 (3)
NICT	0.4320 (3)	0.8799 (4)	0.5537 (4)	0.4301 (4)

NICT	0.4260 (4)	0.8786 (6)	0.5413 (7)	0.4255 (5)
NICT	0.4250 (5)	0.8763 (7)	0.5519 (5)	0.4226 (7)
NICT	0.4250 (5)	0.8798 (5)	0.5481 (6)	0.4240 (6)
NICT	0.4210 (6)	0.8748 (8)	0.5399 (8)	0.4192 (8)
NICT	0.4170 (7)	0.8695 (14)	0.5327 (11)	0.4138 (11)
NICT	0.4160 (8)	0.8710 (12)	0.5320 (13)	0.4150 (9)
NICT	0.4160 (8)	0.8734 (9)	0.5388 (9)	0.4147 (10)
NICT	0.4130 (9)	0.8729 (10)	0.5374 (10)	0.4115 (12)
NICT	0.4110 (10)	0.8723 (11)	0.5309 (14)	0.4099 (13)
NICT	0.4100 (11)	0.8696 (13)	0.5327 (12)	0.4088 (14)
NICT	0.3980 (12)	0.8686 (15)	0.5238 (15)	0.3965 (15)
QIAU	0.3460 (13)	0.8600 (16)	0.4737 (16)	0.3438 (16)
UOH	0.1990 (14)	0.8063 (17)	0.1990 (17)	0.1969 (17)
UOH	0.1830 (15)	0.7958 (18)	0.1830 (18)	0.1814 (18)
UOH	0.1070 (16)	0.7379 (19)	0.1070 (19)	0.1059 (19)
QIAU	0.0000 (17)	0.2583 (20)	0.0000 (20)	0.0000 (20)

Table 9: Results for the English to Bangla transliteration task (EnBa) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.1891 (1)	0.8031 (1)	0.2679 (3)	0.1877 (1)
NICT	0.1882 (2)	0.8023 (3)	0.2695 (2)	0.1868 (2)
NICT	0.1845 (3)	0.8029 (2)	0.2700 (1)	0.1832 (3)
NICT	0.1809 (4)	0.7952 (5)	0.2516 (5)	0.1802 (4)
NICT	0.1800 (5)	0.7951 (7)	0.2506 (6)	0.1793 (5)
NICT	0.1800 (5)	0.7902 (11)	0.2577 (4)	0.1784 (6)
NICT	0.1755 (6)	0.7926 (8)	0.2448 (11)	0.1739 (7)
NICT	0.1745 (7)	0.7896 (14)	0.2460 (9)	0.1731 (8)
NICT	0.1745 (7)	0.7888 (15)	0.2453 (10)	0.1730 (9)
NICT	0.1736 (8)	0.7918 (10)	0.2447 (12)	0.1720 (10)
NICT	0.1727 (9)	0.7902 (12)	0.2485 (7)	0.1711 (11)
NICT	0.1718 (10)	0.7879 (16)	0.2469 (8)	0.1702 (12)
NICT	0.1709 (11)	0.7919 (9)	0.2443 (13)	0.1693 (13)
NICT	0.1709 (11)	0.7831 (17)	0.2404 (15)	0.1693 (13)
NICT	0.1682 (12)	0.7897 (13)	0.2432 (14)	0.1664 (14)
QIAU	0.1591 (13)	0.7976 (4)	0.2377 (16)	0.1582 (15)
UOH	0.1482 (14)	0.7817 (19)	0.1482 (17)	0.1466 (16)
UOH	0.1445 (15)	0.7951 (6)	0.1445 (18)	0.1436 (17)
UOH	0.1382 (16)	0.7828 (18)	0.1382 (19)	0.1366 (18)
UOH	0.1355 (17)	0.7746 (20)	0.1355 (20)	0.1339 (19)

Table 10: Results for the English to Hebrew transliteration task (EnHe) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.3524 (1)	0.7067 (1)	0.4509 (2)	0.3526 (1)
NICT	0.3505 (2)	0.7032 (2)	0.4518 (1)	0.3507 (2)
NICT	0.3476 (3)	0.7022 (3)	0.4505 (3)	0.3476 (3)
NICT	0.3229 (4)	0.6885 (4)	0.4090 (4)	0.3226 (4)
NICT	0.3181 (5)	0.6762 (7)	0.4057 (5)	0.3190 (5)

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.3133 (6)	0.6794 (5)	0.4023 (7)	0.3136 (6)
NICT	0.3124 (7)	0.6735 (10)	0.4042 (6)	0.3126 (7)
NICT	0.3067 (8)	0.6782 (6)	0.3998 (8)	0.3064 (9)
NICT	0.3067 (8)	0.6747 (9)	0.3987 (9)	0.3064 (9)
NICT	0.3067 (8)	0.6659 (12)	0.3946 (11)	0.3069 (8)
NICT	0.3000 (9)	0.6647 (13)	0.3875 (13)	0.2998 (10)
NICT	0.2971 (10)	0.6702 (11)	0.3912 (12)	0.2974 (12)
NICT	0.2971 (10)	0.6624 (15)	0.3815 (14)	0.2979 (11)
NICT	0.2952 (11)	0.6749 (8)	0.3958 (10)	0.2955 (13)
NICT	0.2895 (12)	0.6631 (14)	0.3794 (15)	0.2890 (14)

Table 11: Results for the English to Korean transliteration task (EnKo) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.4647 (1)	0.8386 (1)	0.5916 (1)	0.4637 (1)
NICT	0.4637 (2)	0.8386 (2)	0.5902 (2)	0.4627 (2)
NICT	0.4608 (3)	0.8359 (3)	0.5830 (3)	0.4601 (3)
NICT	0.4182 (4)	0.8160 (5)	0.5434 (4)	0.4177 (4)
NICT	0.4153 (5)	0.8210 (4)	0.5400 (5)	0.4148 (5)
NICT	0.4105 (6)	0.8141 (8)	0.5346 (6)	0.4090 (6)
NICT	0.4066 (7)	0.8155 (6)	0.5259 (9)	0.4056 (7)
NICT	0.4008 (8)	0.8154 (7)	0.5252 (10)	0.3998 (8)
NICT	0.3988 (9)	0.8110 (9)	0.5143 (12)	0.3969 (9)
NICT	0.3979 (10)	0.8099 (11)	0.5270 (8)	0.3964 (11)
NICT	0.3979 (10)	0.8102 (10)	0.5340 (7)	0.3967 (10)
NICT	0.3950 (11)	0.8043 (13)	0.5210 (11)	0.3921 (12)
NICT	0.3843 (12)	0.8029 (14)	0.5129 (13)	0.3824 (13)
NICT	0.3775 (13)	0.8067 (12)	0.4982 (14)	0.3768 (14)
NICT	0.0000 (14)	0.0000 (15)	0.0000 (15)	0.0000 (15)

Table 12: Results for the English to Japanese (Kantakana) transliteration task (EnJa) on Evaluation Test.

Team	Accuracy	F-score	MRR	MAP$_{ref}$
NICT	0.3269 (1)	0.5886 (1)	0.3883 (3)	0.2465 (3)
NICT	0.3251 (2)	0.5867 (2)	0.3911 (1)	0.2489 (2)
NICT	0.3233 (3)	0.5845 (3)	0.3900 (2)	0.2490 (1)
NICT	0.3169 (4)	0.5782 (4)	0.3821 (4)	0.2459 (4)
NICT	0.3160 (5)	0.5698 (5)	0.3724 (6)	0.2385 (5)
NICT	0.3096 (6)	0.5651 (7)	0.3774 (5)	0.2375 (6)
NICT	0.3087 (7)	0.5625 (9)	0.3684 (8)	0.2333 (7)
NICT	0.3041 (8)	0.5545 (12)	0.3689 (7)	0.2263 (8)
NICT	0.3023 (9)	0.5665 (6)	0.3577 (11)	0.2257 (9)
NICT	0.2986 (10)	0.5626 (8)	0.3585 (10)	0.2250 (10)
NICT	0.2950 (11)	0.5562 (10)	0.3606 (9)	0.2249 (11)
NICT	0.2922 (12)	0.5561 (11)	0.3532 (12)	0.2198 (13)
NICT	0.2895 (13)	0.5520 (13)	0.3511 (13)	0.2209 (12)
NICT	0.2868 (14)	0.5506 (14)	0.3474 (14)	0.2195 (14)
NICT	0.2804 (15)	0.5417 (15)	0.3396 (15)	0.2120 (15)

Table 13: Results for the English to Japanese (Kanji) transliteration task (JnJk) on Evaluation Test.