English to Indian Languages Machine Transliteration System at NEWS 2010

Amitava Das[1], **Tanik Saikh**[2], **Tapabrata Mondal**[3], **Asif Ekbal**[4], **Sivaji Bandyopadhyay**[5]
Department of Computer Science and Engineering[1,2,3,5]
Jadavpur University,
Kolkata-700032, India
amitava.santu@gmail.com[1], tanik4u@gmail.com[2], tapabratamondal@gmail.com[3], sivaji_cse_ju@yahoo.com[5]
Department of Computational Linguistics[4]
University of Heidelberg
Im Neuenheimer Feld 325
69120 Heidelberg, Germany
ekbal@cl.uni-heidelberg.de

Abstract

This paper reports about our work in the NEWS 2010 Shared Task on Transliteration Generation held as part of ACL 2010. One standard run and two non-standard runs were submitted for English to Hindi and Bengali transliteration while one standard and one non-standard run were submitted for Kannada and Tamil. The transliteration systems are based on Orthographic rules and Phoneme based technology. The system has been trained on the NEWS 2010 Shared Task on Transliteration Generation datasets. For the standard run, the system demonstrated mean F-Score values of 0.818 for Bengali, 0.714 for Hindi, 0.663 for Kannada and 0.563 for Tamil. The reported mean F-Score values of non-standard runs are 0.845 and 0.875 for Bengali non-standard run-1 and 2, 0.752 and 0.739 for Hindi non-standard run-1 and 2, 0.662 for Kannada non-standard run-1 and 0.760 for Tamil non-standard run-1. Non-Standard Run-2 for Bengali has achieved the highest score among all the submitted runs. Hindi Non-Standard Run-1 and Run-2 runs are ranked as the 5th and 6th among all submitted Runs.

1 Introduction

Transliteration is the method of translating one source language word into another target language by expressing and preserving the original pronunciation in their source language. Thus, the central problem in transliteration is predicting the pronunciation of the original word. Transliteration between two languages that use the same set of alphabets is trivial: the word is left as it is. However, for languages those use different alphabet sets the names must be transliterated or rendered in the target language alphabets. Transliteration of words is necessary in many applications, such as machine translation, corpus alignment, cross-language Information Retrieval, information extraction and automatic lexicon acquisition. In the literature, a number of transliteration algorithms are available involving English (Li et al., 2004; Vigra and Khudanpur, 2003; Goto et al., 2003), European languages (Marino et al., 2005) and some of the Asian languages, namely Chinese (Li et al., 2004; Vigra and Khudanpur, 2003), Japanese (Goto et al., 2003; Knight and Graehl, 1998), Korean (Jung et al., 2000) and Arabic (Al-Onaizan and Knight, 2002a; Al-Onaizan and Knight, 2002c). Recently, some works have been initiated involving Indian languages (Ekbal et al., 2006; Ekbal et al., 2007; Surana and Singh, 2008). The detailed report of our participation in NEWS 2009 could be found in (Das et al., 2009).

One standard run for Bengali (Bengali Standard Run: BSR), Hindi (Hindi Standard Run: HSR), Kannada (Kannada Standard Run: KSR) and Tamil (Tamil Standard Run: TSR) were submitted. Two non-standard runs for English to Hindi (Hindi Non-Standard Run 1 & 2: HNSR1 & HNSR2) and Bengali (Bengali Non-Standard Run 1 & 2: BNSR1 & BNSR1) transliteration were submitted. Only one non-standard run were submitted for Kannada (Kannada Non-Standard Run-1: KNSR1) and Tamil (Tamil Non-Standard Run-1: TNSR1).

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 71–75,
Uppsala, Sweden, 16 July 2010.

2 Machine Transliteration Systems

Five different transliteration models have been proposed in the present report that can generate the transliteration in Indian language from an English word. The transliteration models are named as Trigram Model (Tri), Joint Source-Channel Model (JSC), Modified Joint Source-Channel Model (MJSC), Improved Modified Joint Source-Channel Model (IMJSC) and International Phonetic Alphabet Based Model (IPA). Among all the models the first four are categorized as orthographic model and the last one i.e. IPA based model is categorized as phoneme based model.

An English word is divided into Transliteration Units (TUs) with patterns C*V*, where C represents a consonant and V represents a vowel. The targeted words in Indian languages are divided into TUs with patterns C+M?, where C represents a consonant or a vowel or a conjunct and M represents the vowel modifier or matra. The TUs are the basic lexical units for machine transliteration. The system considers the English and Indian languages contextual information in the form of collocated TUs simultaneously to calculate the plausibility of transliteration from each English TU to various Indian languages candidate TUs and chooses the one with maximum probability. The system learns the mappings automatically from the bilingual NEWS 2010 training set being guided by linguistic features/knowledge. The output of the mapping process is a decision-list classifier with collocated TUs in the source language and their equivalent TUs in collocation in the target language along with the probability of each decision obtained from the training set. A Direct example base has been maintained that contains the bilingual training examples that do not result in the equal number of TUs in both the source and target sides during alignment. The Direct example base is checked first during machine transliteration of the input English word. If no match is obtained, the system uses direct orthographic mapping by identifying the equivalent TU in Indian languages for each English TU in the input and then placing the target language TUs in order. The IPA based model has been used for English dictionary words. Words which are not present in the dictionary are handled by other orthographic models as Trigram, JSC, MJSC and IMJSC.

The transliteration models are described below in which S and T denotes the source and the target words respectively:

3 Orthographic Transliteration models

The orthographic models work on the idea of TUs from both source and target languages. The orthographic models used in the present system are described below. For transliteration, P(*T*), i.e., the probability of transliteration in the target language, is calculated from a English-Indian languages bilingual database If, *T* is not found in the dictionary, then a very small value is assigned to P(*T*). These models have been desribed in details in Ekbal et al. (2007).

3.1 Trigram

This is basically the Trigram model where the previous and the next source TUs are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid s_{k-1}, s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

3.2 Joint Source-Channel Model (JSC)

This is essentially the Joint Source-Channel model (Hazhiou et al., 2004) where the previous TUs with reference to the current TUs in both the source (s) and the target sides (t) are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

3.3 Modified Joint Source-Channel Model (MJSC)

In this model, the previous and the next TUs in the source and the previous target TU are considered as the context. This is the Modified Joint Source-Channel model.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1}, s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

3.4 Improved Modified Joint Source-Channel Model (IMJSC)

In this model, the previous two and the next TUs in the source and the previous target TU are considered as the context. This is the Improved Modified Joint Source-Channel model.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid s_{k+1} < s,t >_{k-1}, s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

4 International Phonetic Alphabet (IPA) Model

The NEWS 2010 Shared Task on Transliteration Generation challenge addresses general domain transliteration problem rather than named entity transliteration. Due to large number of dictionary words as reported in Table 1 in NEWS 2010 data set a phoneme based transliteration algorithm has been devised.

	Train	Dev	Test
Bengali	7.77%	5.14%	6.46%
Hindi	27.82%	15.80%	3.7%
Kannada	27.60%	14.63%	4.4%
Tamil	27.87%	17.31%	3.0%

Table 1: Statistics of Dictionary Words

The International Phonetic Alphabet (IPA) is a system of representing phonetic notations based primarily on the Latin alphabet and devised by the International Phonetic Association as a standardized representation of the sounds of spoken language. The machine-readable Carnegie Mellon Pronouncing Dictionary [1] has been used as an external resource to capture source language IPA structure. The dictionary contains over 125,000 words and their transcriptions with mappings from words to their pronunciations in the given phoneme set. The current phoneme set contains 39 distinct phonemes. As there is no such parallel IPA dictionary available for Indian languages, English IPA structures have been mapped to TUs in Indian languages during training. An example of such mapping between phonemes and TUs are shown in Table 3, for which the vowels may carry lexical stress as reported in Table 2. This phone set is based on the ARPAbet[2] symbol set developed for speech recognition uses.

Representation	Stress level
0	No
1	Primary
2	Secondary

Table 2: Stress Level on Vowel

A pre-processing module checks whether a targeted source English word is a valid dictionary word or not. The dictionary words are then handled by phoneme based transliteration module.

[1] www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] http://en.wikipedia.org/wiki/Arpabet

Phoneme	Example	Translation	TUs
AA	odd	AA0-D	অ-ড
AH	hut	HH0-AH-T	হা-ট
D	dee	D-IY1	ড-ী

Table 3: Phoneme Map Patterns of English Words and TUs

In the target side we use our TU segregation logic to get phoneme wise transliteration pattern. We present this problem as a sequence labelling problem, because transliteration pattern changes depending upon the contextual phonemes in source side and TUs in the target side. We use a standard machine learning based sequence labeller Conditional Random Field (CRF)[3] here.

IPA based model increased the performance for Bengali, Hindi and Tamil languages as reported in Section 6. The performance has decreased for Kannada.

5 Ranking

The ranking among the transliterated outputs follow the order reported in Table 4: The ranking decision is based on the experiments as described in (Ekbal et al., 2006) and additionally based on the experiments on NEWS 2010 development dataset.

Word Type	Ranking Order				
	1	2	3	4	5
Dictionary	IPA	IMJSC	MJSC	JSC	Tri
Non-Dictionary	IMJSC	MJSC	JSC	Tri	-

Table 4: Phoneme Patterns of English Words

In BSR, HSR, KSR and TSR the orthographic TU based models such as: IMJSC, MJSC, JSC and Tri have been used only trained by NEWS 2010 dataset. In BNSR1 and HNSR1 all the orthographic models have been trained with additional census dataset as described in Section 6. In case of BNSR2, HNSR2, KNSR1 and TNSR1 the output of the IPA based model has been added with highest priority. As no census data is available for Kannada and Tamil therefore there is only one Non-Standard Run was submitted for these two languages only with the output of IPA based model along with the output of Standard Run.

6 Experimental Results

We have trained our transliteration models using the NEWS 2010 datasets obtained from the NEWS 2010 Machine Transliteration Shared Task (Li et al., 2010). A brief statistics of the

[3] http://crfpp.sourceforge.net

datasets are presented in Table 5. During training, we have split multi-words into collections of single word transliterations. It was observed that the number of tokens in the source and target sides mismatched in various multi-words and these cases were not considered further. Following are some examples:

Paris Charles de Gaulle पेरिस रॉसे चार्ल्स डे ग्यूले

Suven Life Scie ಸುವೆನ್ ಲೈಫ್ ಸೈನ್ಸ್

Delta Air Lines டெல்டா ஏர்லைன்ஸ்

In the training set, some multi-words were partly translated and not transliterated. Such examples were dropped from the training set. In the following example the English word "National" is being translated in the target as "राष्ट्रीय".

Australian National University ऑस्ट्रेलियन **राष्ट्रीय** यूनिवर्सिटी

Set	Number of examples			
	Bng	Hnd	Kn	Tm
Training	11938	9975	7990	7974
Development	992	1974	1968	1987
Test	991	1000	1000	1000

Table 5: Statistics of Dataset

There is less number of known examples in the NEWS 2010 test set from training set. The exact figure is reported in the Table 6.

	Matches with training
Bengali	14.73%
Hindi	0.2%
Kannada	0.0%
Tamil	0.0%

Table 6: Statistics of Dataset

If the outputs of any two transliteration models are same for any word then only one output are provided for that particular word. Evaluation results of the final system are shown in Table 7 for Bengali, Table 8 for Hindi, Table 9 for Kannada and Table 10 for Tamil.

Parameters	Accuracy		
	BSR	BNSR1	BNSR2
Accuracy in top-1	0.232	0.369	0.430
Mean F-score	0.818	0.845	0.875
Mean Reciprocal Rank (MRR)	0.325	0.451	0.526
Mean Average Precision (MAP)$_{ref}$	0.232	0.369	0.430

Table 7: Results on Bengali Test Set

Parameters	Accuracy		
	HSR	HNSR1	HNSR2
Accuracy in top-1	0.150	0.254	0.170
Mean F-score	0.714	0.752	0.739
Mean Reciprocal Rank (MRR)	0.308	0.369	0.314
Mean Average Precision (MAP)$_{ref}$	0.150	0.254	0.170

Table 8: Results on Hindi Test Set

Parameters	Accuracy	
	KSR	KNSR1
Accuracy in top-1	0.056	0.055
Mean F-score	0.663	0.662
Mean Reciprocal Rank (MRR)	0.112	0.169
Mean Average Precision (MAP)$_{ref}$	0.056	0.055

Table 9: Results on Kannada Test Set

Parameters	Accuracy	
	TSR	TNSR1
Accuracy in top-1	0.013	0.082
Mean F-score	0.563	0.760
Mean Reciprocal Rank (MRR)	0.121	0.142
Mean Average Precision (MAP)$_{ref}$	0.013	0.082

Table 10: Results on Tamil Test Set

The additional dataset used for the non-standard runs is mainly the census data consisting of only Indian person names that have been collected from the web[4]. In the BNSR1 and HNSR1 we have used an English-Bengali/Hindi bilingual census example dataset. English-Hindi set consist of 961,890 examples and English-Bengali set consist of 582984 examples. This database contains the frequency of the corresponding English-Bengali/Hindi name pair.

7 Conclusion

This paper reports about our works as part of the NEWS 2010 Shared Task on Transliteration Generation. We have used both the orthographic and phoneme based transliteration modules for the present task. As our all previous efforts was for named entity transliteration. The Transliteration Generation challenge addresses general domain transliteration problem rather than named entity transliteration. To handle general transliteration problem we proposed a IPA based methodology.

[4]http://www.eci.gov.in/DevForum/Fullname.asp

References

A. Das, A. Ekbal, Tapabrata Mondal and S. Bandyopadhyay. English to Hindi Machine Transliteration at NEWS 2009. In Proceedings of the NEWS 2009, In Proceeding of ACL-IJCNLP 2009, August 7th, 2009, Singapore.

Al-Onaizan, Y. and Knight, K. 2002a. Named Entity Translation: Extended Abstract. In Proceedings of the Human Language Technology Conference, 122– 124.

Al-Onaizan, Y. and Knight, K. 2002b. Translating Named Entities using Monolingual and Bilingual Resources. In Proceedings of the 40th Annual Meeting of the ACL, 400–408, USA.

Ekbal, A. Naskar, S. and Bandyopadhyay, S. 2007. Named Entity Transliteration. International Journal of Computer Processing of Oriental Languages (IJCPOL), Volume (20:4), 289-310, World Scientific Publishing Company, Singapore.

Ekbal, A., Naskar, S. and Bandyopadhyay, S. 2006. A Modified Joint Source Channel Model for Transliteration. In Proceedings of the COLING-ACL 2006, 191-198, Australia.

Goto, I., Kato, N., Uratani, N. and Ehara, T. 2003. Transliteration Considering Context Information based on the Maximum Entropy Method. In Proceeding of the MT-Summit IX, 125–132, New Orleans, USA.

Jung, Sung Young , Sung Lim Hong and Eunok Paek. 2000. An English to Korean Transliteration Model of Extended Markov Window. In Proceedings of International Conference on Computational Linguistics (COLING 2000), 383-389.

Knight, K. and Graehl, J. 1998. Machine Transliteration, Computational Linguistics, Volume (24:4), 599–612.

Kumaran, A. and Tobias Kellner. 2007. A generic framework for machine transliteration. In Proc. of the 30th SIGIR.

Li, Haizhou, A Kumaran, Min Zhang and Vladimir Pervouchine. 2010. Whitepaper: NEWS 2010 Shared Task on Transliteration Generation. In the ACL 2010 Named Entities Workshop (NEWS-2010), Uppsala, Sweden, Association for Computational Linguistics, July 2010.

Li, Haizhou, Min Zhang and Su Jian. 2004. A Joint Source-Channel Model for Machine Transliteration. In Proceedings of the 42nd Annual Meeting of the ACL, 159-166. Spain.

Marino, J. B., R. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa and M. Ruiz. 2005. Bilingual n-gram Statistical Machine Translation. In Proceedings of the MT-Summit X, 275–282.

Surana, Harshit, and Singh, Anil Kumar. 2008. A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), 64-71, India.

Vigra, Paola and Khudanpur, S. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition, 57–60.