# A Modified Joint Source-Channel Model for NE Transliteration

Asif Ekbal
Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
and
EMMA Post-doctoral Student, Department of Computational Linguistics, University of Heidelberg, Germany
Email: ekbal@cl.uni-heidelberg.de asif.ekbal@gmail.com

# Machine Transliteration and Joint Source-Channel Model

- Transliteration System
  - Input: Character string in the source language
  - Output: Character in the target language as output
  - Two Steps of Transliteration
    - Segmentation of the source string into transliteration units (TUs)
    - Relating the source language TUs to the corresponding units in the target language; resolving different combinations of alignments and unit mappings
  - Mathematical Formulation
    - Source language name: S
    - Target language name: T
    - Maximize P(T | S)
    - Bayes' Rule (Source to Target Language Transliteration, S2T):
    - $$P(S,T) = P(S \mid T) \times P(T) \qquad (1)$$

# Machine Transliteration and Joint Source-Channel Model

- $P(S|T)$ → Probability of transliterating T to S through a noisy channel (Transformation rules)
- $P(T)$ → Probability distribution of source
  - Reflects what is considered good target language transliteration in general

- Back Transliteration: Target to Source Transliteration (T2S)
  - (2)

$$P(S,T) = P(T \mid S) \times P(S)$$

  - P(S) and P(T) of (1) and (2) → Estimated using n-gram language models

- Estimation of $P(S|T)$ and $P(T|S)$ using Phoneme-based approach
  - Approximate probability distribution by introducing a phonemic representation
  - Source name S converted into an intermediate phonemic representation P
  - P further converted into the target language name T

# Machine Transliteration and Joint Source-Channel Model

- S2T transliteration
  - $P(T \mid S) = P(T \mid P) \times P(P \mid S)$ (3)
- T2S transliteration
  - $P(S \mid T) = P(S \mid P) \times P(P \mid T)$ (4)
- Joint Source-Channel Model (Hazhiou et al., 2004)
  - Alternative to Phoneme-based approach
  - Based on the close coupling of the source and target transliteration units (TUs)
  - For K aligned TUs

$$
\begin{aligned}
P(S,T) &= P(s_1, s_2, \ldots, s_k, t_1, t_2, \ldots, t_k) \\
&= P(\langle s,t\rangle_1, \langle s,t\rangle_2, \ldots, \langle s,t\rangle_k) \\
&= \prod_{k=1}^{K} P(\langle s,t\rangle_k \mid \langle s,t\rangle_1^{k-1})
\end{aligned}
\qquad (5)
$$

# Machine Transliteration and Joint Source-Channel Model

- Let us consider
  - Source name: $\alpha = x_1x_{2 \ldots\ldots\ldots\ldots} x_m$ [xi, i = 1: m are source TUs]
  - Target name: $\beta = y_1y_{2 \ldots\ldots\ldots} y_n$ [$y_j$, j = 1: n are target TUs]
  - $m \neq n$ (very often) (i.e., Target TU may correspond to one or more Source TUs)
  - Alignment (γ)= <s, t>$_1$ = <$x_1$, $y_1$>; <s, t>$_2$ = <$x_2x_3$, $y_2$>; …….. <s, t>$_k$ = <$x_m$, $y_n$>
  - TU correspondence <s, t> → Transliteration pair
- S2T transliteration →

$$\overline{\beta} = \underset{\beta,\gamma}{\arg\max}\, P(\alpha, \beta, \gamma)$$

# Machine Transliteration and Joint Source-Channel Model

- T2S transliteration → $$\bar{\alpha} = \underset{\alpha,\gamma}{\arg\max}\, P(\alpha, \beta, \gamma)$$

- *n*-gram transliteration model: Conditional probability or transliteration probability of a transliteration pair <s, t>$_k$ depending on its immediate *n* predecessor pairs

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

# Bengali to English Machine Transliteration

- Bengali and English names divided into Transliteration Units (TUs)

  - Regular expression for Bengali TU: $C^{+}M$ ?

    where, C represents a vowel or a consonant or a conjunct and M represents the vowel modifier or matra

  - Regular expression for English TU: C*V*

    where, C represents a consonant and V represents a vowel

- Contextual information in the form of collocated TUs considered

# Bengali to English Machine Transliteration

Examples of TUs :

শচীন (*sachin*) ⟶ [শ | চী | ন]

sachin → [sa | chi | n]

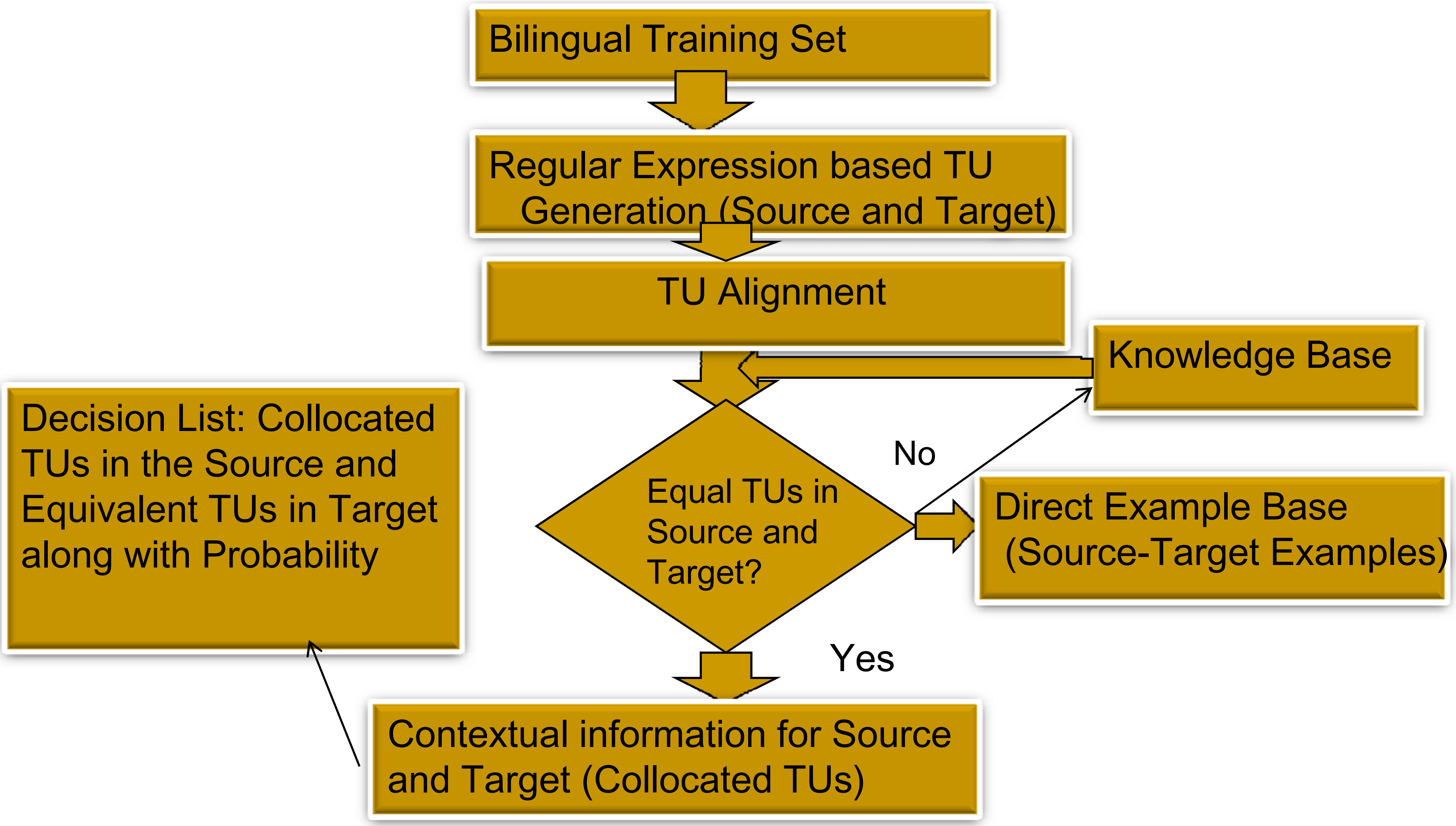
মনোজ (*manoj*)⟶ [ম | নো | জ ]

manoj→ [ ma | no | j ]

শ্রীকান্ত (*srikant*) ⟶ [ শ্রী | কা | ন্ত ]

srikant → [ sri | ka | nt ]

# Overall Procedure

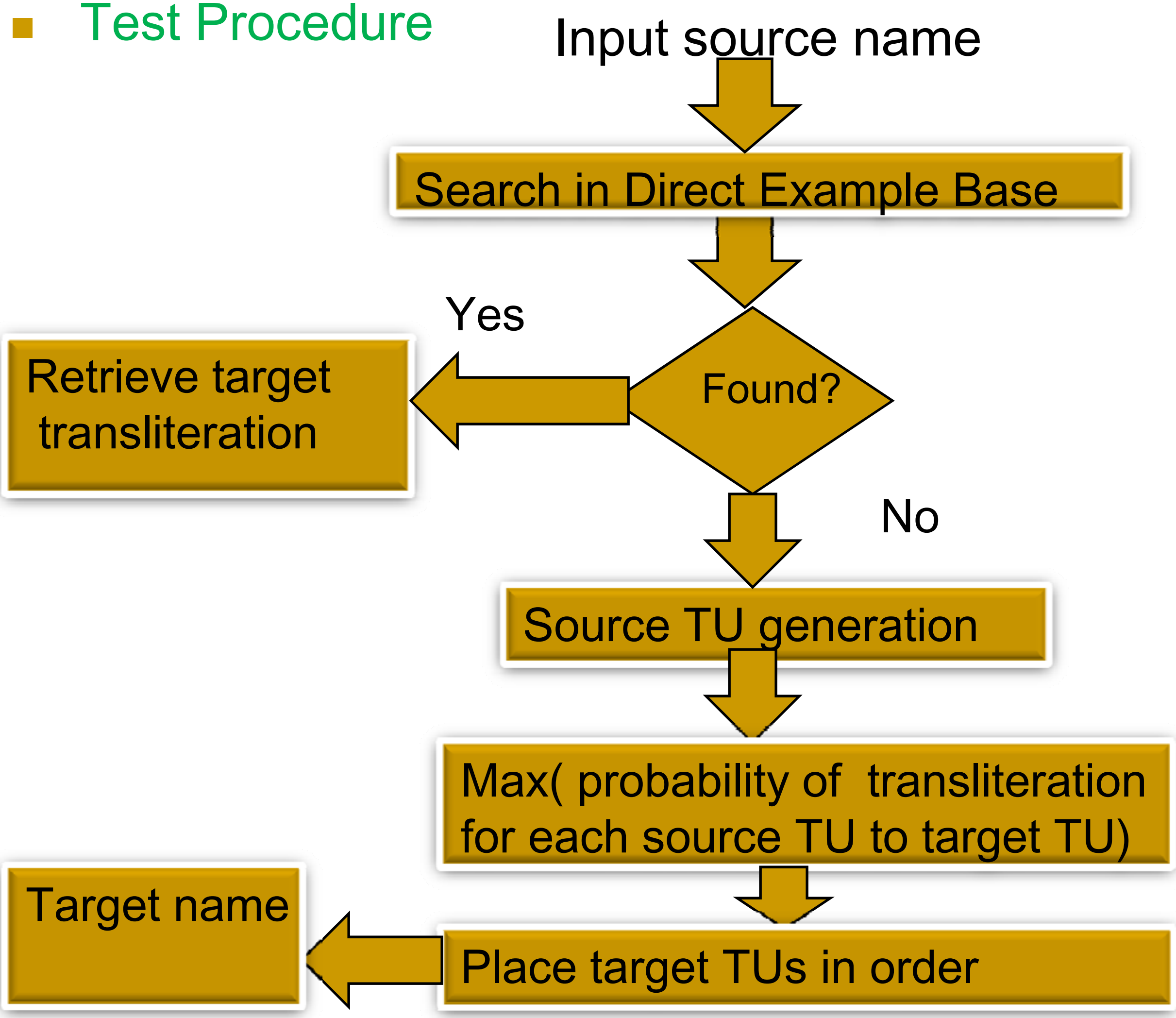


# Overall Procedure (Contd..)

- Bilingual training set: Bengali-English name pairs
- TU Generation: TUs generated according to corresponding regular expression
- TU alignment: Process of mapping each source TU to the target TU
- Number of TUs in the source and target may not be equal
  - Direct Example base: Examples that do not result in one to one correspondence → Language Independent Version
  - Knowledge base: Conjuncts and/or diphthongs in Bengali and their equivalent representations in English → Language Dependent Version
- Output of alignment: Decision-list classifier
  - Collocated TUs in the source language and their equivalent TUs in collocation in the target language
  - Probability of each decision obtained from the training set

# Overall Procedure (Contd..)

- Test Procedure



# Overall Procedure (Contd..)

- Calculate plausibility of transliteration from each source to various target candidate
  - Choose Target candidate TU with maximum probability
  - Appropriate sense of a word in the source language to identify its representation in the target language

- Direct orthographic mapping for transliteration
  - Identify equivalent target TU for each source TU
  - Place Target TUs in order

# Proposed Models for Transliteration

- Baseline Model
  - English consonant / sequence of consonants →Bengali consonant / conjunct/ sequence of consonants
  - English vowels →Bengali vowels/ matra (vowel modifier)
  - English diphthongs→ Vowel/semi-vowel-matra combination in Bengali

- Model A (Monogram): No context in source and target

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

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

- Model B (Bigram): Previous source TU (TU occurring to the left of current TU) as the context

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

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

# Proposed Models for Transliteration (Contd..)

- Model C: Bigram model with next source TU as the context

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

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

- Model D (Joint Source-Channel model) : Previous TUs in source and target as the context

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

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

# Proposed Models for Transliteration (Contd..)

- Model E (Trigram model) :Previous and next source TUs as the context

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

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

- Model F (Modified Joint Source-Channel Model): Previous and the next TUs in the source and the previous target TU as the context

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

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

# Bengali to English Transliteration

- Retrieve TUs from Bengali-English name pair
- Associate the Bengali TUs to the respective English TUs along with the TUs in context
- An Example: রবীন্দ্রনাথ (*rabIndranAth)* → rabindranath

| Source Language | | | Target Language | |
|---|---|---|---|---|
| **Previous TU** | **TU** | **Next TU** | **Previous TU** | **TU** |
| - | র | বী | - | r |
| র | বী | ন্দ্র | bi | r |
| বী | ন্দ্র | না | bi | ndra |
| ন্দ্র | না | থ | ndra | th |
| না | থ | - | th | - |

# Bengali to English Transliteration (Contd..)

- **Problem** : Unequal number of TUs in Source and Target

Example 1: বৃ | জ | মো | হ | ন (*brijmohan*) ↔ bri | jmo | ha | n

Example 2: রা | ই | মা (*raima*) ↔ rai | ma

- Solution:
  - Knowledge base: Lists of Bengali conjuncts and diphthongs and their possible representations in English
- Hypothesis:
  - The problem TU in the English side has always the maximum length

# Bengali to English Transliteration (Contd..)

- Example 1:

  - Same length TUs: *bri* and *jmo*
  - Consult with knowledge
    - Valid conjunct: *bri*
    - Invalid conjunct: *jmo*
    - Split *jmo*
      - Jmo→ *j / mo*
    - New alignment of TUs

      [ বৃ | জ | মো | হ | ন ↔ bri | j | mo | ha | n]

- Example 2:
  - Longest TU in English side: *rai*
  - TU resolved to: ra | i
  - Help of diphthongs

# Bengali to English Transliteration (Contd..)

- Intermediate form of the name pair

  বা | ই | মা (*raima*) ↔ r | ai | ma]

- Matra associated with the Bengali TU that corresponds to English TU *r*
  - A vowel must be attached with TU *r*

- Final TU alignment

  রা | ই | মা (*raima*) ↔ ra | i | ma

.

# Bengali to English Transliteration (Contd..)

- Solution of Knowledge base is not always sufficient
  - Example :

    দে | ব | রা | জ (*devraj*) ↔ de | vra | j

    - Longest TU in English side →vra
    - vr→ Valid conjunct
    - Realignment using knowledge base

    দে | ব | রা | জ (*devraj*) ↔ de | vr | a | j →Wrong alignment

  - Contain constituent Bengali consonants in order and not the conjunct representation
    - Option 1: Remove the conjunct (vr) from the knowledge base

      Put the examples in the *Direct Example Base*
    - Option 2: Do not exclude conjunct from the knowledge base

      Move training examples with constituent consonant representations to the *Direct Example Base*

- Actual realignment : দে | ব | রা | জ (*devraj*) ↔ de | v | ra | j

# Bengali to English Transliteration (Contd..)

- Source and Target TUs may not result into one to one correspondence after the use of linguistic knowledge base
- Examples:
  - Zero-to-one relationship [Φ → h]

    আ | ল্লা (*aallA*) ↔ a | lla | h

    মা | ল | দা (*mAldA*) ↔ ma | l | da | h

  - Many-to-one relationship [আ, ই→ i]

    আ | ই | ভি (*aaivi*)↔ i | vy

    আ | ই | জ | ল (*aaijal*) ↔ i | zwa | l

  - One-to-zero relationship [X → Φ]

    কৃ | ষ্ণ | ন | গ | র *(krishnanagar)* →kri | shna | ga | r

- Step: Put such examples in the Direct Example Base

# Bengali to English Transliteration (Contd..)

- Linguistic knowledge apparently solves mapping problem sometimes
  - Example 1: ব | র | খা ↔ ba | rkha
  - Example 2: ঝা | ড় | খ | ন্ড ↔ jha | rkha | nd

✓ Applying linguistic knowledge (rk→ valid conjunct )

rkha → rk | ha (Example 1 and Example 2)

ব | র | খা ↔ ba | rk | ha (Incorrect TU pair)

ঝা | ড় | খ | ন্ড ↔ jha | rk | ha | nd (Incorrect TU pair)

✓ Actual TU alignment:

ব | র | খা↔ ba | r | kha

ঝা | ড় | খ | ন্ড↔jha | r | kha | nd

- Step: Put such examples in the Direct Example Base

# Evaluation Scheme

- Evaluation Parameters:
  - Transliteration Unit Agreement Ratio (TUAR) and
  - Word Agreement Ratio (WAR)

- ➪ Input Bengali Word : B
- ➪ Gold standard transliteration of the Bengali word : E
- ➪ System generated transliteration of the all input Bengali words : $E^{/}$
- ➪ Err: Total no. of wrongly transliterated TUs in $E^{/}$
- ➪ $Err^{/}$ : Total no. of erroneous names generated by the system

- TUAR = (L-Err ) / L, L: No. of TUs in all E

- WAR = (S- $Err^{/}$ ) / S, S: Test Sample Size

# Evaluation Results

- Two Versions of each models evaluated
  - Language Independent Version (does not use the knowledge of conjuncts and/or diphthongs)
  - Language Dependent Version (uses the knowledge of conjuncts and/or diphthongs)

- Training Set:
  - 25,000 Bengali-English bilingual database
  - Bengali names extracted from a Bengali news corpus (Ekbal and Bandyopadhyay, 2008a) and their transliterations stored manually
  - Person names=18,500
  - Location names=5000
  - Organization names=1500
- Evaluation procedure
  - 5-fold cross validation
  - Consistent error rates with less than 0.5% deviation for each of the 5-fold cross validation tests
  - Random selection of one of the 5 subsets as the standard open test

# Evaluation Results

- Test set→5000

| Test Set Type | Sample Size (S) | Number of TUs (L) | Average Number of TUs Per Name |
|---|---|---|---|
| Person name | 4100 | 18450 | 5 |
| Location name | 675 | 2598 | 4 |
| Organization name | 225 | 1305 | 6 |

Test set statistics

# Results of Language Independent Evaluation (B2E)

Table 1 : Results with evaluation metrics [Training set: 20,000 and Test set: 4000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 52.7 | 76.8 |
| A | 54.4 | 79.5 |
| B | 62.1 | 84.3 |
| C | 59.6 | 82.2 |
| D | 72.5 | 85.2 |
| E | 75.3 | 87.8 |
| F | 76.9 | 91.6 |

# Results of Language Dependent Evaluation (B2E)

Table 2 : Results with evaluation metrics [Training set: 20,000 and Test set: 5000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 52.7 | 76.8 |
| A | 57.8 | 83.3 |
| B | 67.3 | 87.3 |
| C | 64.9 | 85.7 |
| D | 75.8 | 89.8 |
| E | 79.6 | 91.4 |
| F | 81.4 | 95.7 |

# Effects of Linguistic Knowledge during B2E Transliteration

Table 2A: Results with evaluation metrics [Training set: 20,000 and Test set: 5000]

| | With Linguistic Knowledge | | Without Linguistic Knowledge | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR(in %) |
| Baseline | 52.7 | 76.8 | 52.7 | 76.8 |
| A | 57.8 | 83.3 | 54.4 | 79.5 |
| B | 67.3 | 87.3 | 62.1 | 84.3 |
| C | 64.9 | 85.7 | 59.6 | 82.2 |
| D | 75.8 | 89.8 | 72.5 | 85.2 |
| E | 79.6 | 91.4 | 75.3 | 87.8 |
| F | 81.4 | 95.7 | 76.9 | 91.6 |

# Results of Language Independent Evaluation (E2B)

Table 3 : Results with evaluation metrics [Training set: 20,000 and Test set: 5000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 51.8 | 76.6 |
| A | 53.5 | 79.4 |
| B | 61.4 | 82.5 |
| C | 59.5 | 81.9 |
| D | 73.4 | 84.6 |
| E | 73.8 | 87.2 |
| F | 74.8 | 89.6 |

# Results of Language Dependent Evaluation (E2B)

Table 4 : Results with evaluation metrics [Training set: 4,000 and Test set: 5000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 51.8 | 76.6 |
| A | 56.4 | 83.2 |
| B | 65.4 | 85.5 |
| C | 62.6 | 83.6 |
| D | 76.7 | 89.3 |
| E | 77.4 | 91.5 |
| F | 79.5 | 93.8 |

# Effects of Linguistic Knowledge during E2B Transliteration

Table 4A: Results with evaluation metrics [Training set: 20,000 and Test set: 5000]

| | Without linguistic knowledge | | With linguistic knowledge | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR (in %) |
| Baseline | 51.8 | 76.6 | 51.8 | 76.6 |
| A | 53.5 | 79.4 | 56.4 | 83.2 |
| B | 61.4 | 82.5 | 65.4 | 85.5 |
| C | 59.5 | 81.9 | 62.6 | 83.6 |
| D | 73.4 | 84.6 | 76.7 | 89.3 |
| E | 73.8 | 87.2 | 77.4 | 91.5 |
| F | 74.8 | 89.6 | 79.5 | 93.8 |

# Results of Language Independent Evaluation (B2E)

- 5000 bilingual examples randomly selected from the 25000 bilingual examples
  - Training set → 4000 out of 5000 bilingual examples
  - Test set → 1000 out of 5000 bilingual examples

Table 5 : Results with evaluation metrics [Training set: 4,000 and Test set: 1000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 47.1 | 71.3 |
| A | 47.2 | 75.3 |
| B | 54.9 | 79.6 |
| C | 54.6 | 78.1 |
| D | 58.9 | 80.2 |
| E | 62.4 | 83.3 |
| F | 66.3 | 86.5 |

# Effects of Linguistic Knowledge during B2E Transliteration

Table 5A: Results with evaluation metrics [Training set: 4,000 and Test set: 1000]

| | Without Linguistic Knowledge | | With Linguistic Knowledge | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR(in %) |
| Baseline | 47.1 | 71.3 | 47.1 | 71.3 |
| A | 47.2 | 75.3 | 49.3 | 77.2 |
| B | 54.9 | 79.6 | 58.2 | 81.6 |
| C | 54.6 | 78.1 | 56.8 | 80.7 |
| D | 58.9 | 80.2 | 60.8 | 82.2 |
| E | 62.4 | 83.3 | 65.7 | 86.4 |
| F | 66.3 | 86.5 | 69.8 | 89.6 |

# Effects of Data Size during B2E Transliteration

| | Training =4000, Test=1000 | | Training =20000, Test=5000 | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR(in %) |
| Baseline | 47.1 | 71.3 | 52.7 | 76.8 |
| A | 49.3 | 77.2 | 57.8 | 83.3 |
| B | 58.2 | 81.6 | 67.3 | 87.3 |
| C | 56.8 | 80.7 | 64.9 | 85.7 |
| D | 60.8 | 82.2 | 75.8 | 89.8 |
| E | 65.7 | 86.4 | 79.6 | 91.4 |
| F | 69.8 | 89.6 | 81.4 | 95.7 |

# Results of Language Independent Evaluation (E2B)

Table 6 : Results with evaluation metrics [Training set: 4000 and Test set: 1000]

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| Baseline | 45.9 | 70.2 |
| A | 45.4 | 74.9 |
| B | 50.6 | 76.5 |
| C | 48.6 | 75.9 |
| D | 57.6 | 77.6 |
| E | 61.9 | 81.8 |
| F | 65.7 | 85.5 |

# Effects of Linguistic Knowledge during E2B Transliteration

Table 6A: Results with evaluation metrics [Training set: 4000 and Test set: 1000]

| | Without linguistic knowledge | | With linguistic knowledge | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR (in %) |
| Baseline | 45.9 | 70.2 | 45.9 | 70.2 |
| A | 45.4 | 74.9 | 47.2 | 76.3 |
| B | 50.6 | 76.5 | 52.5 | 79.3 |
| C | 48.6 | 75.9 | 51.6 | 78.5 |
| D | 57.6 | 77.6 | 60.5 | 81.7 |
| E | 61.9 | 81.8 | 64.3 | 84.1 |
| F | 65.7 | 85.5 | 67.9 | 87.5 |

# Effects of Data Size during E2B Transliteration

| | Training =4000, Test=1000 | | Training =20000, Test=5000 | |
|---|---|---|---|---|
| Model | WAR (in %) | TUAR (in %) | WAR (in %) | TUAR(in %) |
| Baseline | 45.9 | 70.2 | 52.7 | 76.8 |
| A | 47.2 | 76.3 | 57.8 | 83.3 |
| B | 52.5 | 79.3 | 67.3 | 87.3 |
| C | 51.6 | 78.5 | 64.9 | 85.7 |
| D | 60.5 | 81.7 | 75.8 | 89.8 |
| E | 64.3 | 84.1 | 79.6 | 91.4 |
| F | 67.9 | 87.5 | 81.4 | 95.7 |

# Results for Hindi to English Transliteration

- Training Set: Created from the 4000 Bengali-English examples with the help of GIST SDK toolkit (http://www.cdac.in/html/gist/down/sdk_d.asp)
- Some manual corrections required after the font conversions

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| A | 45.3 | 73.8 |
| B | 54.4 | 78.4 |
| C | 52.6 | 77.3 |
| D | 56.3 | 80.2 |
| E | 61.4 | 81.7 |
| F | 64.8 | 85.7 |

# Results for Telugu to English Transliteration

- Training Set: Created from the 4000 Bengali-English examples with the help of GIST SDK toolkit (http://www.cdac.in/html/gist/down/sdk_d.asp)
- Some manual corrections

| Model | WAR (in %) | TUAR (in %) |
|---|---|---|
| A | 42.7 | 71.8 |
| B | 51.7 | 75.3 |
| C | 49.7 | 74.9 |
| D | 54.6 | 78.2 |
| E | 59.2 | 79.7 |
| F | 62.2 | 82.4 |

# Conclusion

- Modified Joint Source-Channel Model (Model F) performs best in all the cases
- Linguistic knowledge helps to improve system performance
- Most of the errors are t the *matra* level, i.e., a short *matra* might have been replaced by a long *matra* or vice versa
- More linguistic knowledge is necessary to disambiguate the short and the long vowels and the *matra* representations in Bengali
- Inclusion of triphthongs and tetraphthongs
- TU alignment process is general and applicable for the pair of languages that share a comparable orthography

# Relevant Publications

1. A. Ekbal, S. Naskar and S. Bandyopadhyay (2007). Named Entity Transliteration. International Journal of **Computer Processing of Oriental Languages** (**IJCPOL**), Vol. 20(4), 289-310, World Scientific Press, Singapore.

2. **A. Ekbal**, S. Naskar and S. Bandyopadhyay (2007). Language Independent Named Entity Transliteration. In *Proceedings of 3rd Indian International Conference on Artificial Intelligence, Natural Language Independent Engineering Track,* India, PP: 1936-1950.

3. A. Ekbal, S. Naskar and S. Bandyopadhyay (2006). A Modified Joint Source-Channel Model for Transliteration. In *Proceedings of COLING/ACL 2006*, Sydney, Australia, pp. 191-198.