English-Hindi Transliteration using Multiple Similarity Metrics

Niraj Aswani and Robert Gaizauskas

Department of Computer Science, University of Sheffield

English-Hindi Transliteration Mappings

English	Hindi	English	Hindi	English	Hindi
a	ए, अ, ॅ , ा , आ, ऐ, ऍ, ॅ , ॅ , ॲ, ऐ	m, mm	म, ं	h	ः , ह, ऐच
aa	ा , आ	n	ं , न, ऩ, ण, ँ	hh	ह
ae	ए, अ, ॅ , आ, ऐ, ऍ, ॅ , ॅ , ॲ, ऐ	ng	ङ	i	ई, ि, इ, ी , ाय , आय
aei	ऐ, ए, ऐ, ऍ	nn	ण, न, ऩ	ii	ई
ai	ॅ	nu	न	j, jj	ज, ज़
au	औ, ौ	ny	ञ	k, kk	क, क़
aum	ॐ	o	ऑ, ओ, आ, ॊ , ओ, अ, ॉ , ॊ , ा	kh	ख, ख़
b	ब, ब	oau	ॊ , औ	l, ll	ळ, ऌ, ऴ, ॡ, ल
bh	भ	om	ॐ	t	ट
c	स, क, सी, क़	on	न	t, tt	त, ट
cc	स, क, क़	oo	ू , ऊ	th	थ, ठ
ch	क, च, क़	ou	ओ, ौ	u	उ, ि, ऊ, ी , ु , ू
chh	छ	oum	ॐ	v, vv, w	व
d	ज, ड, द, ड़	p	प, पी	y, yy	य, य़, ी
dd	द, ड, ड़	ph	फ, फ़	z, zz	ज, झ, ज़
dh	ढ, ध, ढ़	pp	प	er	यर
e	ऐ, ॅ , ॅ , ि, ऐ, इ, ए, ऍ, ॅ , ॅ , ॅ	qu	क	f	एफ
ea	ॅ , ॅ , ॅ , ए, ऐ	r	र, ् , ऋ, ऱ, ॅ	n	एन
ee	ि, ी , ई	roo	ऋ, ऋ	g, gg	ग, ज, ज़, ग
ei	ॅ	s	स, ॅ , ज, ाँ , ज़, ाँ	gh	घ
es	ॅ , ाँ , ाँ	sh, ss, ti, tio	ष, स, श	rr	र, ऋ, र
f, ff	फ, फ़	ss	ज, ज़	ru	ऋ, ॅ, ॅ ऋ,

cancer vs कैंसर (cancer)

Obtain mappings for each letter of the Hindi word कैंसर

Hindi Letter	English Letters
क	c, k and ch
ै	e, a and ai
ं	n
स	c and s
र	r

Remove mappings which do not appear in the target word

Hindi Letter	English Letters
क	c, k and ~~ch~~
ै	e, a and ~~ai~~
ं	n
स	c and ~~s~~
र	r

Cencr Kencr
Chencr Cancr
Kancr Chancr
Caincr Kaincr
Chaincr Censr
Kensr Chensr
Cansr Kansr
Chansr Cainsr
Kainsr Chainsr

18 candidate transliterations to 2 candidate transliteration

Cencr ● ~~Kencr~~
~~Chencr~~ Cancr ●
~~Kancr~~ ~~Chancr~~
~~Caincr~~ ~~Kaincr~~
~~Chaincr~~ ~~Censr~~
~~Kensr~~ ~~Chensr~~
~~Cansr~~ ~~Kansr~~
~~Chansr~~ ~~Cainsr~~
~~Kainsr~~ ~~Chainsr~~

how vs होमवर्क (homework)

Replacing no mapping with underscore

Hindi Letter	English Letters
ह	h
ो	o
म	m
व	v and w
र्	r
क	c, k and ch

Hindi Letter	English Letters
ह	h
ो	o
म	~~m~~ – replacing with underscore
व	~~v~~ and w
र्	~~r~~ – replacing with underscore
क	~~c~~, ~~k~~ and ~~ch~~ – replacing with underscore

homvrc homwrc
homvrk homwrk
homvrch homwrch

6 candidate transliterations to only 1 candidate transliteration

ho_w__

Transliteration Similarity Metric (TSM)

```
READ S, T   //source and target strings

//initialize variables
SET i=1, j=1, n=|S|, m=|T|, matches=0

// if the shorter string is at least 65% of the length of the longer string
IF (|S|/|T|) >= 0.65 || (|S|/|T|) >= 0.65 THEN

  // check start and end constraints
  // one of the first two  or the last two characters of the two strings must match
  IF (S[1] == T[1] || S[2] == T[1] || S[1] == T[2]) &
     (S[n] == T[m] || S[n-1] == T[m] || S[n] == T[m-1]) THEN

     WHILE i <= n & j <= m       // comparing characters one by one
       FOR k = j to j+2          // matching within the window of 2 characters
         IF S[i] == T[k] THEN

           // found a match
           INCREMENT matches, i
           SET j to k + 1
           CONTINUE WHILE
         ENDIF
       ENDFOR

       INCREMENT i          // character at position i in S does not exist in T
     ENDWHILE
  ENDIF
ENDIF

COMPUTE sim = matches*2/(|S|+|T|)  //computing similarity
RETURN sim  // return the computed similarity
```

Computing similarity for the examples explained above

cancer vs cancr = 5 matches * 2 / total length 11 = 0.90 TSM sim
cencer vs cancr = 4 matches * 2 / total length 11 = 0.72 TSM sim
Homework vs ho_w__ = 2 matches * 2 / total length 14 = 0.29 TSM sim

Evaluation

Test Data: 1000 randomly collected unique pairs from the EMILLE corpus with 732 correct and 268 incorrect transliteration pairs.

Task: Check performance of different similarity metrics.

Performance of Individual Similarity Metrics

Dice's Coefficient Metric

Threshold >=	Precision	Recall	F-measure
95	0.97	0.26	0.41
90	0.93	0.45	0.61
85	0.9	0.59	0.71
80	0.82	0.73	0.78
75	0.77	0.73	0.75
70	0.73	0.73	0.73

Transliteration Similarity Metric

Threshold >=	Precision	Recall	F-measure
95	0.99	0.17	0.29
90	0.97	0.33	0.49
85	0.95	0.51	0.67
80	0.92	0.61	0.73
75	0.81	0.73	0.77
70	0.73	0.73	0.73

LCSR Metric`

Threshold >=	Precision	Recall	F-measure
95	0.92	0.32	0.48
90	0.92	0.34	0.5
85	0.92	0.41	0.56
80	0.87	0.48	0.62
75	0.86	0.51	0.64
70	0.85	0.55	0.67

N-gram (3) Metric

Threshold >=	Precision	Recall	F-measure
95	0.99	0.15	0.27
90	0.99	0.15	0.27
85	0.99	0.16	0.27
80	0.99	0.17	0.29
75	0.99	0.2	0.34
70	0.99	0.24	0.39

Jaro-Winkler Metric

Threshold >=	Precision	Recall	F-measure
95	0.99	0.18	0.31
90	0.95	0.33	0.49
85	0.93	0.46	0.61
80	0.86	0.58	0.69
75	0.83	0.62	0.71
70	0.8	0.68	0.73

Lavenshtein Distance Metric

Threshold >=	Precision	Recall	F-measure
95	0.99	0.15	0.27
90	1	0.19	0.32
85	0.97	0.32	0.48
80	0.93	0.47	0.62
75	0.9	0.58	0.7
70	0.88	0.64	0.74

Multiple Measure Agreement Strategy*

Why? Given the different criteria that these similarity metrics work on, it is possible that given a pair of strings one metric gives it a very high score whereas the others very low.

How? Top combination of metrics that performed best (f-measure) given different threshold values.

Results

Group	Threshold	F-Measure
DS + TSM + JW	>= 75	0.85
DS + TSM + JW	>= 78	0.86
DS + TSM + JW	**>= 79**	**0.92**
DS + TSM + JW	**>= 80**	**0.92**
DS + TSM + JW	>= 81	0.91
DS + TSM + JW	>= 85	0.78
DC + TSM + LCSR	>= 90	0.62
DC + LCSR + JW	>= 95	0.4

Combination of Dice's Coeffient, Jaro-Winkler and the TSM metric works best with threshold value set between 0.79 and 0.81

Test Data: 2500 English-Hindi sentence pairs.

Task: Compare each noun word in the source English sentence with every word in the Hindi target sentence. Ask the three methods DC, TSM and JW to cast their votes. If a word pair receives at least two votes the pair is considered as a transliteration pair.

Result: 1078 pairs found with 94.71% accuracy.

Experiments with the Gujarati Language

The scripts used by the Gujarati and the Hindi languages have similar consonants and vowels that are pronounced the same way. We replaced the Hindi letters in our mappings with their corresponding Gujarati letters.

0	૦	a	એ, અ, ા, આ, ઍ, ે, ૅ	c	સ, સી, ક	f	ફ, એફ
1	૧	aa	ા, આ	cc	સ, ક	ff	ફ
2	૨	ae	એ, ઍ, ે	ch	ક, ચ	g	ગ, જ, જ્ઞ
3	૩	aum, om, oum	ૐ	chh	છ	gg	ગ, જ
4	૪	b	બ, બી	ai	ૈ	gh	ઘ
5	૫	o	ઑ, ઓ, આ, ો, અ, ા	dd	દ, ડ,	h	હ, એચ
6	૬	oau	ો, ઔ	dh	ઢ, ધ	hh	હ
7	૭	n	ં, ન, એન, ણ, ઁ	ei	ૈ	ii	ઈ
8	૮	q	ક્યુ, ક્યૂ, ક	k	ક, કે	kh	ખ
9	૯	ru	ુ, ઋ, ૃ	ll	લ	m	મ, એમ, ં
tt	ત, ટ	t	ત, ટ, ટી	nn	ણ, ન	un	ન
kk	ક	w, ww	વ	on	ન	oo	ૂ, ઊ, ુ
mm	મ	d	જ, ડ, દ, ડ, ડી	ph	ફ	pp	પ
l	લ, એલ	e, ee	ે, એ, ૈ, િ, ઍ, ઈ, ઈ, ી	th	થ, ઠ	roo	ઋ
ou	ઔ, ી	ea	ૈ, ે, એ, ઍ	i	ા	sh	ષ, સ, શ, ધ
qu	ક	r	ર, આર, ુ, ઋ, ૃ	vv	વ	ti	શ
rr	ર, ઋ	s	સ, એસ, જ, ોં, ાઁ	y	ય	ue	યુ, યૂ, ુ, ૂ
z, zz	જ, ઝ	u	ઉ, િ, ઊ, ી, ુ, યુ, યૂ, ૂ	er	યર	p	પ, પી
tio	શ, ધ	es	ે, ોં, ાઁ	aei	ઍ, એ	bb	બ
v	વ, વી	i	ઈ, િ, ઇ, ી, આય, આઈ	j	જ, જે	bh	ભ
yy	ય	ss	શ, ષ, ધ, સ, જ	au	ઔ, ી	jj	જ

Test Data: 500 English-Gujarati sentence pairs.

Task: Compare each noun word in the source English sentence with every word in the Hindi target sentence. Ask the three `methods DC, TSM and JW to cast their votes. If a word pair receives at least two votes the pair is considered as a transliteration pair.

Result: 450 word pairs found with 90.7% accuracy.

Conclusion

We proposed a bi-directional character(s) mappings. We presented an algorithm for computing a similarity measure. By evaluating various similarity metrics individually and together under a multiple measure agreement scenario we showed that it is possible to identify English-Hindi word pairs that are translation of each other with fairly high frequency. By adapting our system to the Gujarati language we showed that our system is portable.