Konkanverter

A Finite State Machine based Statistical Machine Transliteration Engine for the Konkani Language

Vinodh Rajan
vrs3@st-andrews.ac.uk
http://www.virtualvinodh.com

University of St Andrews

FOUNDED 1413

PhD Student
Computer Science

Konkani - Language

- Indo-European language spoken by around 2.5 million people across India
- One of the 22 scheduled languages in the 8th schedule of the Indian Constitution
- Official language in the Indian state of Goa
- Significant minority language in the states of Karnataka, Kerala & Maharashtra

University of St Andrews
FOUNDED 1413

Konkani - Script(s?)

- As many as five different scripts are used
- *Devanagari Script* is official in the state of Goa
  - But the Goan catholics also use *Roman Script*
- *Kannada Script* is used in Karnataka
- *Malayalam Script* is used in Kerala
- Some Muslim sections of the community use the *Perso-Arabic Script*

Konkani - Script(s?)

Complex socio-religious factors are involved in the script issue !

Konknni

ಕೊಂಕ್ಣಿ
koṃkṇi

كونكنى
kvnkny

കൊങ്കണി
koṅkaṇi

कोंकणी
kōṃkaṇī

http://m.outlookindia.com/story.aspx/?sid=4&aid=227761

Main Orthographies

Devanagari

Implicit schwa deletion

Idiosyncratic vowel lengths

Character sequences are not confounded

Kannada

Explicit schwa deletion

Idiosyncratic vowel lengths

Character sequences are not confounded

Romi

Explicit schwa deletion

Vowel lengths not differentiated

Character sequences frequently merged

University of St Andrews
FOUNDED 1413

Main Orthographies

Devanagari	Kannada	Romi
देवनागरी dēvanāgarī	ದೇವ್‌ನಾಗರಿ dēvnāgari	devnagori
झटको jhaṭakō	ಝಟ್ಕೊ jhaṭko	zhottko
धैर्य dhairya	ದೈರ್ಯ್ dhairy	dhoirio
देवादयेन dēvādayēna	ದೆವಾದಯೆನ್ devādayen	devadoien
सत्तेवयल्या sattēvayalyā	ಸತ್ತೆವಯ್ಲ್ಯಾ sattevaylyā	sot'tevoilea

Examples

Initial Attempts

- Non-availability of parallel corpus
- Implemented initial rule-based transliteration engine
- Manually mined the transliteration rules
- System was not very accurate
  - Especially complex conversions involving Romi

University of St Andrews
FOUNDED 1413

Corpus Creation

- Konkani is an under-resourced language
- Creation of parallel corpus to move towards statistical methods and improve accuracy
- Arranged workshops, recruited people and collaborated with language scholars to create a sizeable corpus

University of St Andrews
FOUNDED 1413

Corpus Creation

- Available to anyone on request

Orthography	Word Count
Devanagari - Kannada	23 187
Devanagari - Romi	38 550
Kannada - Romi	14 396

University of St Andrews
FOUNDED 1413

Konkanverter

- Transliteration engine to convert between all three major Konkani orthographies
- Consists of cascading finite state transducers
- Incorporates both rule-based and statistical approaches

University of St Andrews
FOUNDED 1413

Architecture

- Konkanverter is primarily built on *OpenFst* framework
- Rule-based transducers were written using *Thrax*
- Character alignment was performed using *Phonetisaurus*
- N-gram models were built with *OpenGrm Ngram Library*

University of St Andrews
FOUNDED 1413

Devanagarí to Kannada

Input Romanization

- Uses a custom romanization scheme
- Convert the script from syllabic form to alphabetic form to ease further processing

तोड़ूंक → tōḍūṁka

Let $\mathcal{J}$ denote the romanized input.

University of St Andrews
FOUNDED 1413

This presentation will use ISO 15919 instead of the custom romanization scheme for readability

Devanagarí to Kannada

Schwa Deletion

- Mark morphemic boundaries
- Create an FST $\mathcal{W}_d$ to apply schwa deletion rules

  (e.g) Deleting schwa at word-final positions etc.

āsāta	→	āsāt
āṃvadō	→	āṃvdō
payasa	→	pays
āḍakātī	→	āḍkātī

Refer to paper for the complete set of rules

Devanagarí to Kannada

Rule-based processing

- Create FSTs for all consistent transliteration rules

  (e.g) Word-finally, Devanagari ī/ū is Kannada i/u

  satī → sati

Overall rule-based transduction $\mathcal{R}_{kna} = \mathcal{W}_d \circ \mathcal{R}_{kn}$

Input $\mathcal{I}$ is composed with the above and its output projection is taken

$$\mathcal{R}_{knp} = \pi_2(\mathcal{R}_{kna} \circ \mathcal{I})$$

Refer to paper for the complete set of rules

University of St Andrews
FOUNDED 1413

Devanagarí to Kannada

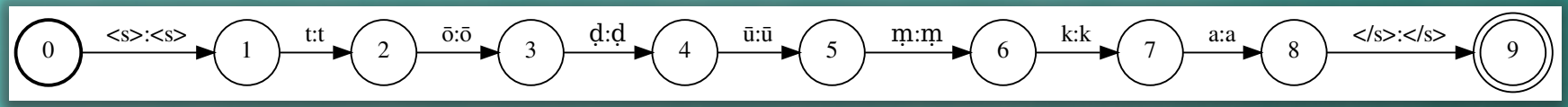
Additional Arcs

- Add additional arcs to $\mathcal{R}_{knp}$ denoting inconsistent rules
  - (e.g) ī/ū before final schwa-consonants with/ without a preceding r/ṃ is either retained or made short.

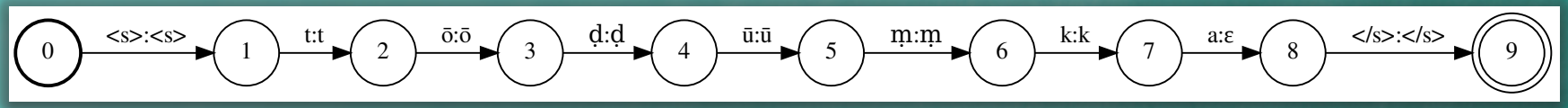
Thus $\mathcal{R}_{knp}$ is modified into $\mathcal{R}_{knm}$

Refer to paper for the complete set of rules

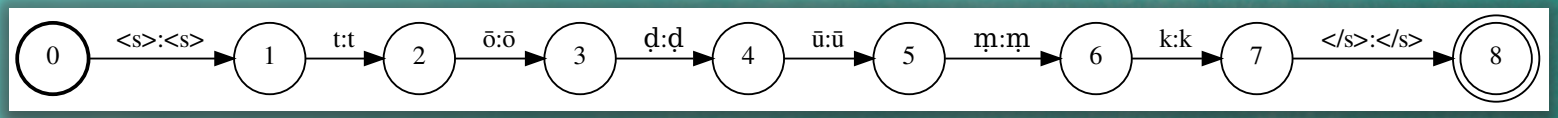
Devanagarí to Kannada



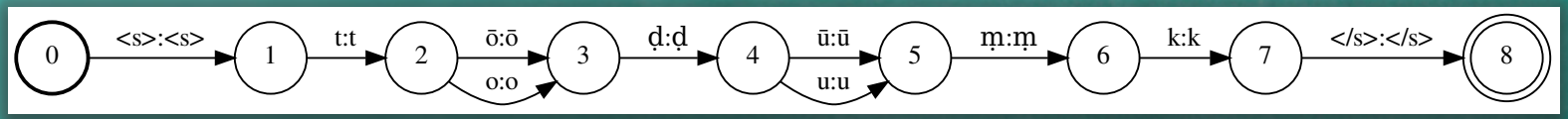
$\mathcal{J}$



$\mathcal{R}_{kna} \circ \mathcal{J}$



$\mathcal{R}_{knp}$



$\mathcal{R}_{knm}$

University of St Andrews
FOUNDED 1413

Devanagarí to Kannada

Lexical Acceptor

- Kannada lexicon created using the words in corpus
- Lexical Acceptor $\mathcal{A}_{\mathcal{L}kn}$ accepts only words in the lexicon

$$\mathcal{T}_{knl} = \mathcal{R}_{knm} \circ \mathcal{A}_{\mathcal{L}kn}$$

Devanagarí to Kannada

N-gram Model

- A character n-gram model $\mathcal{N}_{kn}$ is created from the lexicon
- If all outputs are non-lexical or multiple lexical outputs, the best output is selected based on n-gram probabilities

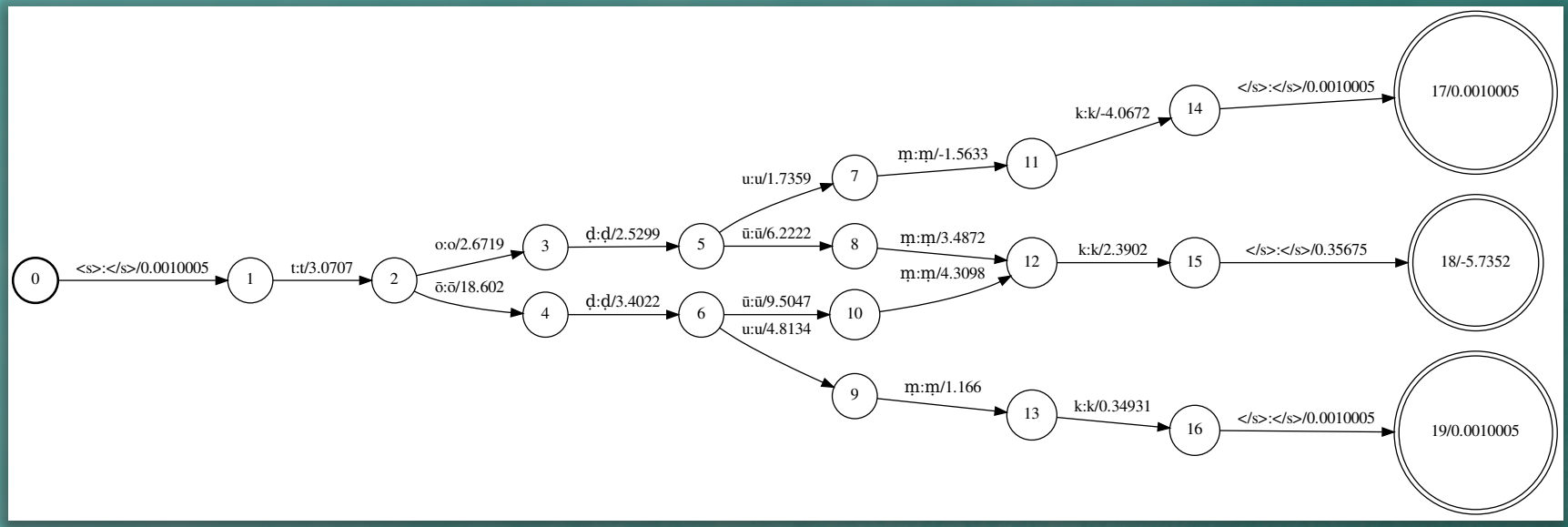
$$\mathcal{T}_{dv2kn} = \mathcal{R}_{knm} \circ \mathcal{N}_{kn}$$

(or)

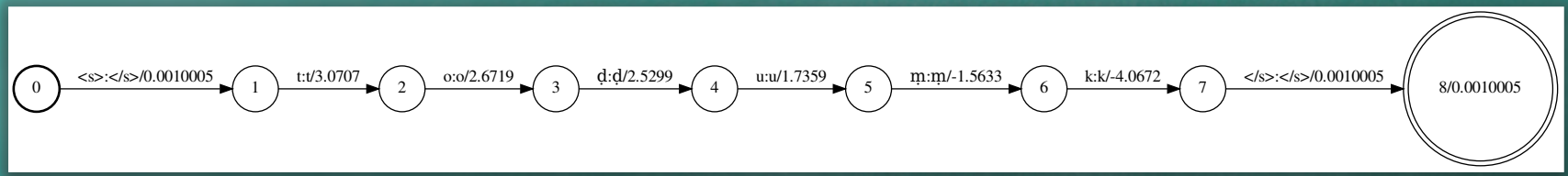
$$\mathcal{T}_{dv2kn} = \mathcal{T}_{knl} \circ \mathcal{N}_{kn}$$

University of St Andrews
FOUNDED 1413

Devanagarí to Kannada



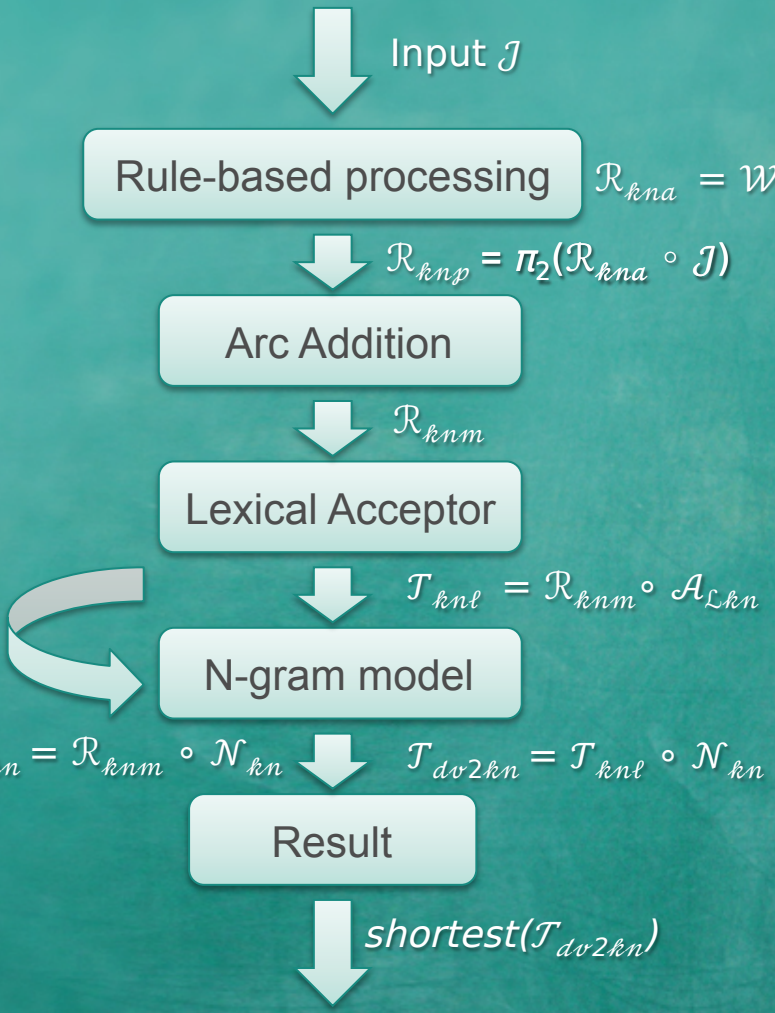
$\mathcal{T}_{dv2kn}$



$shortest(\mathcal{T}_{dv2kn})$

University of St Andrews
FOUNDED 1413

Devanagarí to Kannada



$\mathcal{R}_{kna} = \mathcal{W}_d \circ \mathcal{R}_{kn}$

Summary

Kannada to Devanagarí

Schwa Insertion

- Invert $\mathcal{W}_d$ to insert Schwa
  - But this creates multiple alternate outputs
  - Create a Cluster Acceptor ($\mathcal{A}_{Cdv}$) that rejects all non-standard clusters

Kannada cikṭun → *Devanagari* cikṭūna, cikaṭūna

Kannada vastu → *Devanagari* vastū, vasatū

University of St Andrews
FOUNDED 1413

Kannada to Devanagarí

Rule-based Processing

- Invert rules created in Devanagari to Kannada
- Add necessary additional rules
  - Kannada e/o corresponds to Devanagari ē/ō

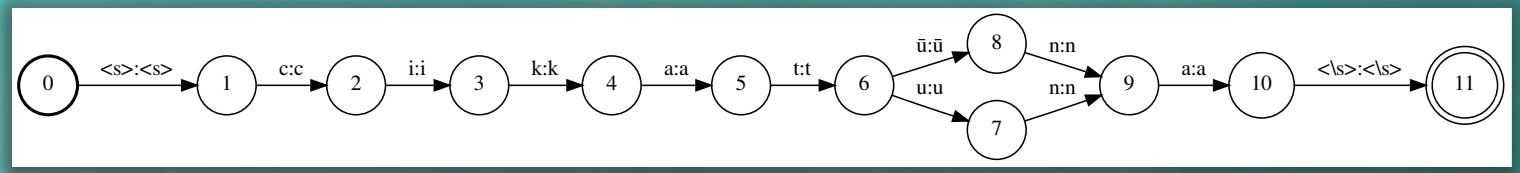
$$\mathcal{R}_{dv} = \mathcal{R}_{kn}^{-1} \circ (\mathcal{W}_{d}^{-1} \circ \mathcal{A}_{Cdv}) \circ \mathcal{R}_{eo}$$

- If composition fails with $\mathcal{W}_{d}^{-1}$

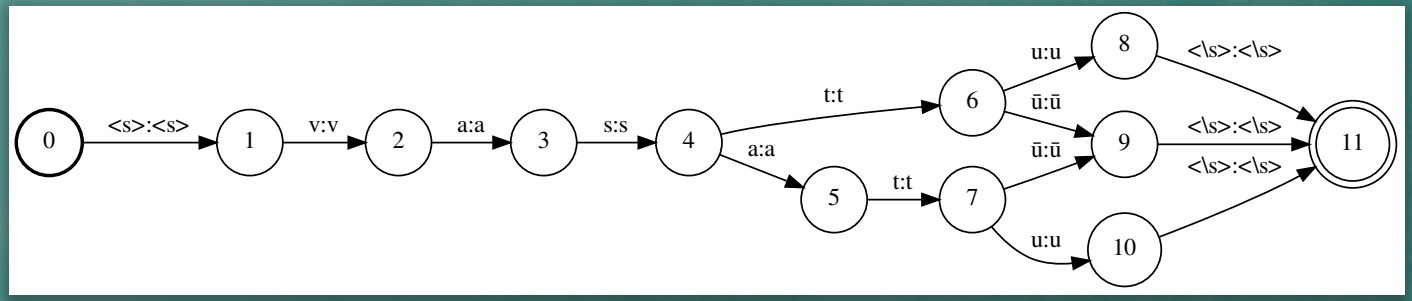
$$\mathcal{R}_{dv} = \mathcal{R}_{kn}^{-1} \circ \mathcal{W}_{ir} \circ \mathcal{R}_{eo}$$

University of St Andrews
FOUNDED 1413

Kannada to Devanagarí



$\mathcal{R}_{dv}$ for cikṭun



$\mathcal{R}_{dv}$ for vastu

University of St Andrews
FOUNDED 1413

Kannada to Devanagarí

Lexical Acceptor & N-Gram Model

- Similar to Devanagari-to-Kannada processing, we use a lexical acceptor and an n-gram model to choose the most-probable output

$$\mathcal{T}_{kn2dv} = (\pi_2(\mathcal{R}_{dv} \circ \mathcal{J}) \circ \mathcal{A}_{\mathcal{L}dv}) \circ \mathcal{N}_{dv}$$

(or)

$$\mathcal{T}_{kn2dv} = \pi_2(\mathcal{R}_{dv} \circ \mathcal{J}) \circ \mathcal{N}_{dv}$$

University of St Andrews
FOUNDED 1413

Kannada to Romi

- Kannada & Romi use different graphemic sets which don't have a one-to-one correspondence
- We use a different approach to transliterate between these scripts that is more statistical than the Devanagari-Romi pair

University of St Andrews
FOUNDED 1413

Kannada to Romi

Joint Sequence N-gram Model

- Align Kannada-Romi wordlist at the character level

meḷilleṃ | mellil'lem → m}m e}e l}l|l i}i l}l|' l}l e}e ṃ}m

gaḍyeṃtlyān | gaddientlean → g}g ā}a ḍ}d|d y}i e}e ṃ}n t}t l}l y}e ā}a n}n

- Create a joint sequence n-gram model $\mathcal{N}_{knrm}$ based on the aligned corpus
- Compose the romanized input with the n-gram model

University of St Andrews
FOUNDED 1413

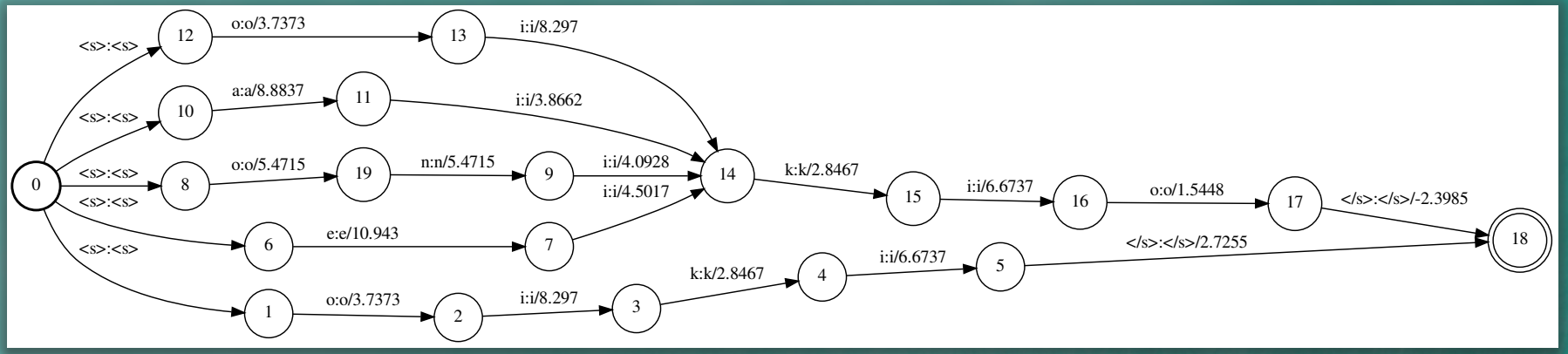
Kannada to Romí

Prune Lattice

- Prune the lattice by rejecting non-standard paths such as those containing geminated vowel digraphs
- Let this transducer be $\mathcal{A}_{rm}$ and the pruned lattice be

$$\mathcal{L}_p = \pi_2(\mathcal{J} \circ \mathcal{N}_{knrm}) \circ \mathcal{A}_{rm}$$

University of St Andrews
FOUNDED 1413



$\mathcal{L}_p$ for Input *aiky* (5-best paths)

University of St Andrews
FOUNDED 1413

Kannada to Romi

Lexical Acceptor & N-gram model

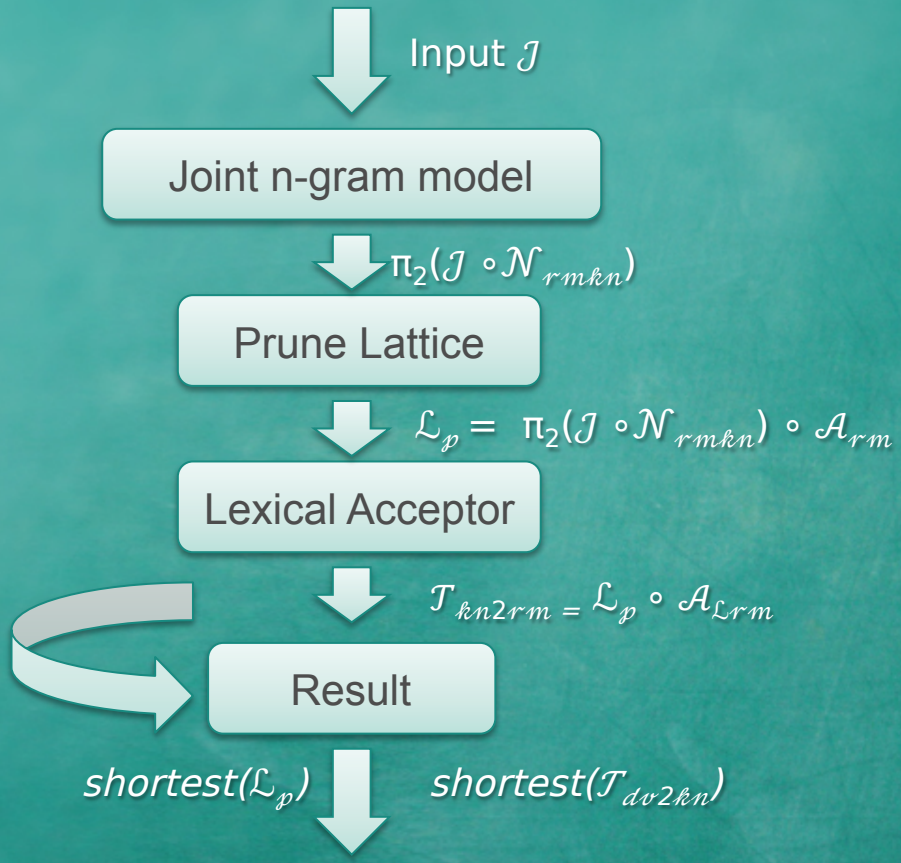
- The pruned lattice is then composed with a lexical acceptor $\mathcal{A}_{\mathcal{L}rm}$
- If the lexical acceptor rejects all paths, choose the optimal paths based on n-gram probabilities

$$\mathcal{T}_{kn2rm} = \mathcal{L}_p \circ \mathcal{A}_{\mathcal{L}rm}$$

(or)

$$\mathcal{T}_{kn2rm} = shortest(\mathcal{L}_p)$$

Kannada to Romi



University of St Andrews
FOUNDED 1413

Summary

Romi to Kannada

- Use a similar method as Kannada to Romi
- But train a new joint sequence n-gram model with Romi-Kannada parallel wordlist
  - Inverting Kannada-Romi n-gram model produces less accurate results

$$\mathcal{L}_p = \pi_2(\mathcal{J} \circ \mathcal{N}_{rmkn}) \circ \mathcal{A}_{ind}$$

(or)

$$\mathcal{T}_{rm2kn} = \mathcal{L}_p \circ \mathcal{A}_{Lkn}$$

$$\mathcal{T}_{rm2kn} = shortest(\mathcal{L}_p)$$

University of St Andrews
FOUNDED 1413

Devanagarí to Romí

Pre-processing (Schwa Deletion)

- Pre-process the parallel corpus by deleting schwa using $\mathcal{W}_d$
- Pre-processing removes one of the underlying grammatical uncertainties
  - Improves the character alignment, resulting in a better n-gram model

University of St Andrews
FOUNDED 1413

Devanagarí to Romí

- Perform a similar set of transductions as in Kannada to Romi
- The input is also schwa deleted

$$\mathcal{L}_p = \pi_2((\mathcal{J} \circ \mathcal{W}_d) \circ \mathcal{N}_{dvrm}) \circ \mathcal{A}_{rm}$$

(or)

$$\mathcal{T}_{dv2rm} = \mathcal{L}_p \circ \mathcal{A}_{\mathcal{L}rm}$$

$$\mathcal{T}_{dv2rm} = shortest(\mathcal{L}_p)$$

University of St Andrews
FOUNDED 1413

Romi to Devanagari

Pre-processing (Schwa Insertion)

- We perform schwa insertion $\mathcal{W}_{ir}$ on the input
  - Not very effective due to the confounding of several original Indic sequences with vowel digraphs/trigraphs

University of St Andrews
FOUNDED 1413

Romí to Devanagarí

- Perform a similar set of transductions as in Romi to Kannada
- The input also has schwa inserted.

$$\mathcal{L}_p = \pi_2((\mathcal{J} \circ \mathcal{W}_{ir}) \circ \mathcal{N}_{rmdv}) \circ \mathcal{A}_{ind}$$

$$\mathcal{T}_{dv2rm} = \mathcal{L}_p \circ \mathcal{A}_{\mathcal{L}dv}$$

$$\mathcal{T}_{dv2rm} = shortest(\mathcal{L}_p)$$

Evaluation

- The corpus was split into training & test set
  - 90% for training : 10% for testing
- Compare rule-based, statistical and cascading systems
- Pure rule-based system evaluated using the initial system developed
- Pure statistical system evaluated using Phonetisaurus' G2P system

Results

Script Pair	Rules-based System	Statistical System	Cascading System
Devanagari - Kannada	83.9%	84.59%	90.383%
Kannada - Devanagari	79.49%	90.16%	96.66%
Devanagari - Romi	74.88%	78.02%	95.39%
Romi - Devanagari	54.02%	74.04%	83.41%
Kannada - Romi	81.29%	87.63%	96.12%
Romi - Kannada	68.01%	82.21%	97.87%

Results

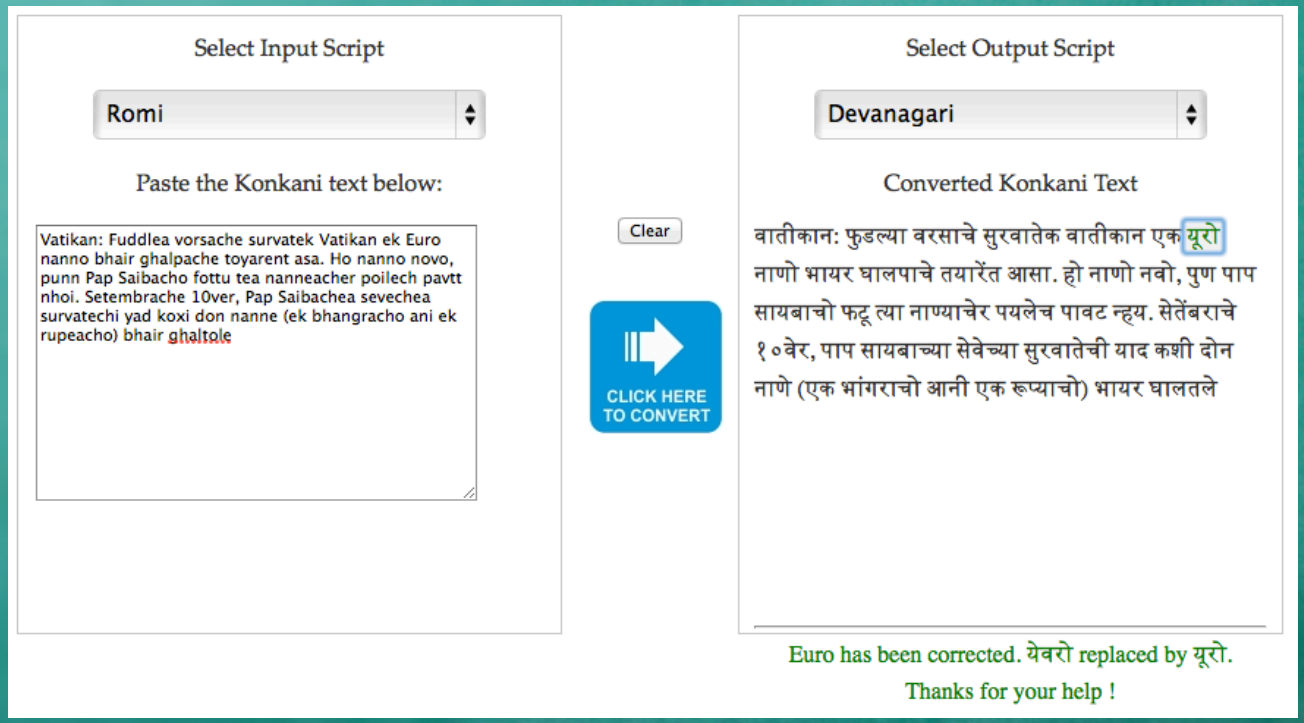
- The rule-based system has the lowest accuracy
  - It is practically impossible to mine all rules
- Statistical system performs at a mediocre level
  - This is due to the limited corpus data
- The cascading approach performs the best with limited corpus

University of St Andrews
FOUNDED 1413

Live System

http://vrs3.host.cs.st-andrews.ac.uk/konkanverter/

- Users can submit corrections to improve accuracy



University of St Andrews
FOUNDED 1413

Future Work

- Research on the minority orthographies
  - Malayalam Script
  - Perso-Arabic Script
- Integrate these minority orthographies into Konkanverter

University of St Andrews
FOUNDED 1413

Acknowledgements

कोंकणी भास आनी संस्कृती प्रतिष्ठान • ಕೊಂಕಣಿ ಭಾಸ ಆನಿ ಸಂಸ್ಕೃತಿ ಪ್ರತಿಷ್ಠಾನ • Konknni Bhas Ani Sounskruti Prothistaan
Konkani Language and Cultural Foundation

University of St Andrews
FOUNDED 1413

Dev Borem Korum !

ദേവ് ബരെം കരും !

ದೇವ್ ಬರೆಂ ಕರುಂ !

देव बरें करूं !

دیو بریں کروں !

Thanks !
&
Questions ?

University of St Andrews
FOUNDED 1413