Translation & Transliteration between Related Languages

Anoop Kunchukuttan
Research Scholar, CFILT, IIT Bombay
anoopk@cse.iitb.ac.in

Mitesh Khapra
Researcher, IBM India Research Lab
mikhapra@in.ibm.com

The Twelfth International Conference on Natural Language Processing (*ICON-2015*)
Thiruvananthapuram, India

11th December 2015

Under the guidance of Prof. Pushpak Bhattacharyya

Download tutorial slides from:
www.cfilt.iitb.ac.in/resources/surveys/icon_2015_tutorial_smt_related_languages.pdf

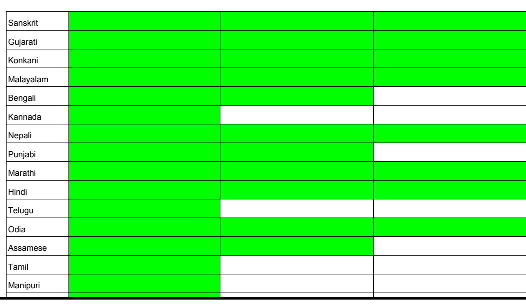
Can you guess the meaning?

ज्ञानम्	परमम्	ध्येयम्
gyanam	*paramam*	*dhyeyam*

Can you guess the meaning?

The synonym *uddeshya* covers more languages

ज्ञानम्	परमम्	ध्येयम्
gyanam	*paramam*	*dhyeyam*
knowledge	supreme	goal



Can you read this?

અમદાવાદ રેલ્વે સ્ટેશન

Can you read this?

અમદાવાદ રેલ્વે સ્ટેશન

अमदावाद रेल्वे स्टेशन

amadAvAda relve sTeshana

- Indic scripts are very similar
- If you learn one, learning others is easy
- Pronunciation of the same word may vary

Tutorial Part 1

- Motivation
- Notions of Language Relatedness
  - Language Families (Genetic)
  - Linguistic Area
  - Language Universals
  - Script
- A Primer to SMT

Tutorial Part 2

- Leveraging Orthographic similarity for transliteration
  - Rule-based transliteration for Indic scripts
  - Akshar-based statistical transliteration for Indic scripts
- Leveraging Lexical Similarity
  - Reduce out-of-vocabulary words & parallel corpus requirements
    - String/Phonetic Similarity
    - Cognate/Transliteration Mining
    - Improve word alignment
    - Transliterating OOV words
  - Character-oriented SMT

Tutorial Part 3

- Leveraging Morphological Similarity
  - Word Segmentation to improve translation
- Leveraging Syntactic Similarity
  - Sharing source reordering rules for translation between two groups of related languages
- Synergy among Multiple Languages
  - Pivot/Bridge languages
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources
- Q&A

Where are we?

- **Motivation**
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

How can relatedness help for translation & transliteration?

Motivation

- Universal translation has proved to be very challenging
- The world is going "glocal" - trends in politics, economics & technology
- Huge translation requirements are between related languages
  - Within a set of related languages
  - Between a *lingua franca* (English, Hindi, Spanish, French, Arabic) and a set of related languages
  - e.g. Indian subcontinent, European Union, South-East Asia
- "Potential" availability of resources between related languages: bilingual speakers, parallel corpora, literature, movies, media
- The unique cultural situation in India - widespread multilingualism

The unique cultural situation in India

- **5+1 language families**
  - Indo-Aryan (74% population)
  - Dravidian (24%)
  - Austro-Asiatic (1.2%)
  - Tibeto-Burman (0.6%)
  - Andaman languages (2 families?)
  - + English (West-Germanic)
- **22 scheduled languages**
- **11 languages with more than 25 million speakers**
  - 29 languages with more than 1 million speakers
  - Only India has 2 languages (+English) in the world's 10 most spoken languages
  - 7-8 Indian languages in the top 20 most spoken languages
- **Greenberg's Linguistic Diversity Index: 0.93**
  - Ranked 9th
  - Highest ranked country outside Pacific Islands and Africa countries
- **The distribution is skewed**: The top 29 languages (>1 million speakers) account for 98.6% of the population
- **125 million English speakers**, highest after the United states

Key similarities between related languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA — *Marathi*

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA — *Marathi segmented*

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA — *Hindi*

- **Lexical:** share significant vocabulary (cognates & loanwords)
- **Morphological:** correspondence between suffixes/post-positions
- **Syntactic:** share the same basic word order

Translating between related languages is easier

Of course, there are differences too ...

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA — *Marathi*

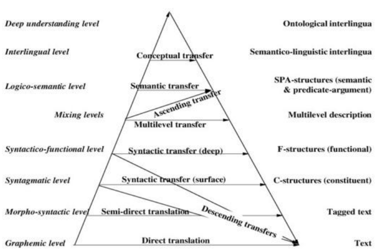
भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA — *Marathi segmented*

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA — *Hindi*

Differences

- Phonetics: *affricative sounds, predominant use of ण (.Na) and ळ (La) in Marathi*
- Morphology: *sandhi rules in Marathi*
- Function words & suffixes:
  a. Hindi uses post-positions, Marathi uses suffixes
  b. Surface forms differ though there are correspondences between Hindi postpositions and Marathi suffixes

Vauquois triangle



- *The central task of MT is bridging language divergence*
- *This task is easier for related languages because:*
  - *Lesser language divergence*
  - *Divergence at lower layers of NLP (for certain types of relatedness)*
  - *More statistical regularities at lower layers of NLP*

A model for translation between close languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA — *Marathi*

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA — *Marathi segmented*

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA — *Hindi*

- Traverse the sentence in sequence one word at a time
- For each word, decide on the action to take:
  - **Transliterate** (Content words primarily)
  - Translate (Function words & suffixes primarily)
  - Skip
  - Insert

- This is a simplified, abstract model
- Monotone decoding

Questions for Discussion

- What does it mean to say languages are related?
- Can translation between related languages be made more accurate?
- Can multiple languages help each other in translation?
- Can we reduce resource requirements?

- Universal translation seems difficult. Can we find the right level of linguistic generalization?
- Can we scale to a group of related languages?

- What concepts and tools are required for solving the above questions?

Where are we?

- Motivation
- **Language Relatedness**
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Relatedness among Languages

Various Notions of Language Relatedness

- Genetic relation → Language Families
- Contact relation → *Sprachbund* (Linguistic Area)
- Linguistic typology → Linguistic Universal
- Orthography → Sharing a script

Genetic Relations

- **Genetic Relations**
- Contact Relations
- Linguistic Typology
- Orthographic Similarity

Language Families

- Group of languages *related through descent from a common ancestor*, called the proto-language of that family

	Sanskrit	Greek	Latin
'father'	*pitā*	*patēr*	*pater*
'foot'	*pad-*	*pod-*	*ped-*
'blood'	*krūra-*	*kreas*	*cruor*
'three'	*trayah*	*treis*	*trēs*
'that'	*tad*	*to*	*-tud*

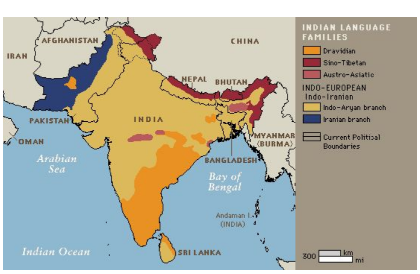
- *Regularity of sound change* is the basis of studying genetic relationships

Meaning	Latin	Portuguese[2]	Castilian	Italian	Romanian
'eight' 'milk' 'fact'	*octo* /ˈoktoː/ *lactem* /ˈlaktẽ/ *factum* /ˈfaktũ/	*oito* /ˈojtu/ *leite* /ˈlɐjtə/ *feito* /ˈfɐjtu/	*ocho* /ˈotʃo/ *leche* /ˈletʃe/ *hecho* /ˈetʃo/	*otto* /ˈɔtto/ *latte* /ˈlatte/ *fatto* /ˈfatto/	*opt* /ˈopt/ *lapte* /ˈlapte/ *fapt* /ˈfapt/

Source: Eifring & Theil (2005)

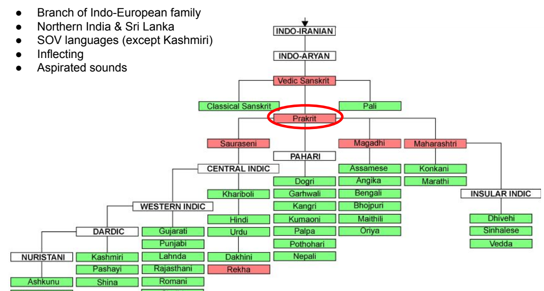
Language Families in India

A study of genetic relations shows 4 major independent language families in India



Indo-Aryan Language Family

- Branch of Indo-European family
- Northern India & Sri Lanka
- SOV languages (except Kashmiri)
- Inflecting
- Aspirated sounds



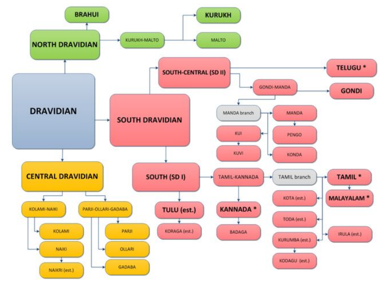
Examples of Cognates

English	Vedic Sanskrit	Hindi	Punjabi	Gujarati	Marathi	Odia	Bengali
bread	rotika	chapātī, roṭī	roṭi	paũ, roṭlā	chapāti, poli, bhākarī	pauruṭi	(pau-)ruṭi
fish	matsya	machhlī	machhī	māchhli	māsa	mācha	machh
hunger	bubuksha, kshudhā	bhūkh	pukh	bhukh	bhūkh	bhoka	khide
language	bhāshā, vāNī	bhāshā, zabān	boli, zabān, pasha	bhāshā	bhāshā	bhāsā	bhasha
ten	dasha	das	das, daha	das	dahā	dasa	dôsh

Source: Wikipedia

Dravidian Languages

- Spoken in South India, Sri Lanka
- SOV languages
- Agglutinative
- Inflecting
- Retroflex sounds



Examples of Cognates

English	Tamil	Malayalam	Kannada	Telugu
fruit	pazham , kanni	pazha.n , phala.n	haNNu , phala	pa.nDu , phala.n
fish	mInn	matsya.n , mIn, mIna.n	mInu , matsya , jalavAsi, mIna	cepalu , matsyalu , jalaba.ndhu
hunger	paci	vishapp , udarArtti , kShutt , pashi	hasivu, hasiv.e,	Akali
language	pAShai, m.ozhi	bhASha , m.ozhi	bhASh.e	bhAShA , paluku
ten	pattu	patt,dasha.m, dashaka.m	hattu	padi

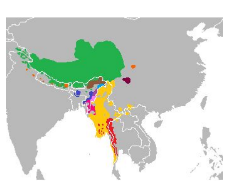
Source: IndoWordNet

Austro-Asiatic Languages

- Austro is *south* in Latin; nothing to to do with languages of Australia
- Munda branch of this family is found in India
  - Ho, Mundari, Santhali, Khasi
- Related to Mon-Khmer branch of S-E Asia: Khmer, Mon, Vietnamese
- Spoken primarily in some parts of Central India (Jharkhand, Chattisgarh, Orissa, WB, Maharashtra)
- From Wikipedia:
  "Linguists traditionally recognize two primary divisions of Austroasiatic: the Mon–Khmer languages of Southeast Asia, Northeast India and the Nicobar Islands, and the Munda languages of East and Central India and parts of Bangladesh. However, no evidence for this classification has ever been published."
- SOV languages
  - exceptions: Khasi
  - They are believed to have been SVO languages in the past (Subbarao, 2012)
- Polysynthetic and Incorporating

Tibeto-Burman language family

- Most spoken in the North-East and the Himalayan areas
- Major languages: Mizo, Meitei, Bodo, Naga, etc.
- Related to Myanmarese, Tibetan and languages of S-E Asia
- SOV word order
- Agglutinative/Isolating depending on the language



What does genetic relatedness imply?

- Cognates (words of the same origin)
- Similar phoneme set, makes transliteration easier
- Similar grammatical properties
  - morphological and word order symmetry makes MT easier
- Cultural similarity leading to shared idioms and multiwords
  - **hi:** दाल में कुछ काला होना (*dAla me.n kuCha kAlA honA*) (something fishy)
    gu: दाळ मा काईक काळु होवु (*dALa mA kAIka kALu hovu*)
  - **mr**: बापाचा माल (*bApAcA mAla*) **hi:** बाप का माल (*bApa kA mAla*)
  - **hi**: वाट लग गई (*vATa laga gaI*) **gu:** वाट लागी गई (*vATa lAgI gaI*) (in trouble)
    mr: वाट लागली (*vATa lAgalI*)
- **Less language divergence leading to easier MT**

Does not necessarily make MT easier
e.g. English & Hindi are divergent in all aspects important to MT viz. lexical, morphological and structural

Language Contact

- Genetic Relations
- **Contact Relations**
- Linguistic Typology
- Orthographic Similarity

- **Linguistic Area**
- Code-Mixing
- Language Shift
- Pidgins & Creoles

Linguistic Area (*Sprachbund)*

- To the layperson, Dravidian & Indo-Aryan languages would seem closer to each other than English & Indo-Aryan
- **Linguistic Area:** A group of languages (at least 3) that have common structural features due to geographical proximity and language contact *(Thomason 2000)*
- Not all features may be shared by all languages in the linguistic area

Examples of linguistic areas:

- **Indian Subcontinent** *(Emeneau, 1956; Subbarao, 2012)*
- Balkans
- South East Asia
- Standard Average European
- Ethiopian highlands
- Sepik River Basin (Papua New Guinea)
- Pacific Northwest

Consequences of language contact

- **Borrowing of vocabulary**

Lexical items are more easily borrowed than grammar and phonology

- **Adoption of features from other languages**
- Stratal influence
- Language shift

Mechanisms for borrowing words *(Eifring & Theil,2005)*

	form	content	example
direct loan	yes	yes	*sushi* < Jap. *sushi*
loanshift	no	yes	*write* (orig. 'draw') < Lat. *scribere*
loan translation	no	yes	*paper tiger* < Ch. *zhǐ lǎohǔ*
loan creation	no	yes	Ch. *diàn-nǎo*, lit. 'electric brain' < *computer*
loanblend	partly	yes	Hindi/Urdu *ḍabal kamrā* < *double room*

- Borrowing phonetic form vs semantic content
- Open class words are more easily borrowed than closed class words
- Nouns are more easily borrowed than verbs
- Peripheral vocabulary is more easily borrowed than basic vocabulary
- Derivational Affixes are easily borrowed

Borrowing of Vocabulary (1)

Sanskrit, Indo-Aryan words in Dravidian languages

- Most classical languages borrow heavily from Sanskrit
- *Anecdotal wisdom*: Malayalam has the highest percentage of Sanskrit origin words, Tamil the lowest

Examples

Sanskrit word	Dravidian Language	Loanword in Dravidian Language	English
cakram	Tamil	cakkaram	wheel
matsyah	Telugu	matsyalu	fish
ashvah	Kannada	ashva	horse
jalam	Malayalam	jala.m	water

Source: IndoWordNet

Borrowing of Vocabulary (2)

Dravidian words in Indo-Aryan languages

- A matter of great debate
- Could probably be of Munda origin also
- See writings of Kuiper, Witzel, Zvelebil, Burrow, etc.
- Proposal of Dravidian borrowing even in early Rg Vedic texts

Borrowing of Vocabulary (3)

- English words in Indian languages
- Indian language words in English
  - Through colonial & modern exchanges as well as ancient trade relations

Examples

- yoga
- guru
- mango
- sugar
- thug
- juggernaut
- cash

Borrowing of Vocabulary (4)

- Words of Persio-Arabic origin

Examples

- khushi
- dIwara
- darvAjA
- dAsTana
- shahara

Vocabulary borrowing - the view from traditional Indian grammar *(Abbi, 2012)*

- Tatsam words: Words from Sanskrit which are used as it is
  - e.g. *hasta*
- Tadbhav words: Words from Sanskrit which undergo phonological changes
  - e.g. *haatha*
- Deshaj words: Words of non-Sanskrit origin in local languages
- Videshaj words: Words of foreign origin *e.g* English, French, Persian, Arabic

Adoption of features in other languages

- Retroflex sounds in Indo-Aryan languages *(Emeneau, 1956; Abbi, 2012)*
  - Sounds: ट ठ ड ढ ण
  - Found in Indo-Aryan, Dravidian and Munda language families
  - Not found in Indo-European languages outside the Indo-Aryan branch
  - But present in the Earliest Vedic literature
  - Probably borrowed from one language family into others a long time ago
- Echo words *(Emeneau, 1956; Subbarao, 2012)*
  - Standard feature in all Dravidian languages
  - Not found in Indo-European languages outside the Indo-Aryan branch
  - Generally means *etcetera* or *things like this*
  - Examples:
    - **hi**: *cAya-vAya*
    - **te:** *pull-gull*
    - **ta** *v.elai-k.elai*

Adoption of features in other languages

Grammar with wide scope is more easily borrowed than grammar with a narrow scope

- SOV word order in Munda languages *(Subbarao, 2012)*
  - Exception: Khasi
  - Their Mon-Khmer cousins have SVO word order
  - Munda language were originally SVO, but have become SOV over time
- Dative subjects *(Abbi, 2012)*
  - Non-agentive subject (generally experiencer)
  - Subject is marked with dative case, and direct object with nominative case
    - **hi**: rAm <u>ko</u> nIndа AyI
    - **ml**: rAm-<u>inna</u> urakkam vannu

Adoption of features in other languages

- Conjunctive participles *(Abbi, 2012; Subbarao, 2012)*
  - used to conjoin two verb phrases in a manner similar to conjunction
  - Two sequential actions; first action expressed with a conjunctive participle
  - **hi**: wah khAnA <u>khAke</u> jAyegA
  - **kn**: mazhA <u>band-u</u> kere tumbitu
rain come tank fill
The tank filled as a result of rain
  - **ml**: mazhA <u>vann-u</u> kuLa.n niranju
rain come pond fill
The pond filled as a result of rain
- Quotative *(Abbi, 2012; Subbarao, 2012)*
  - Reports some one else's quoted speech
  - Present in Dravidian, Munda, Tibeto-Burman and some Indo-Aryan languages (like Marathi, Bengali, Oriya)
  - *iti* (Sanskrit), *asa* (Marathi), *enna* (Malayalam)
  - **mr:** *mi udyA yeto asa to mhNalA*
I tomorrow come +quotative he said

Adoption of features in other languages

- Compound Verb *(Abbi, 2012; Subbarao, 2012)*
  - Verb (Primary) +Verb (vector) combinations
  - Found in very few languages outside Indian subcontinent
  - Examples:
    - **hi**: गिर गया (gira gayA) (fell go)
    - **ml**: വീണു പോയി (viNNu poyI) (fell go)
    - **te: పడి పోయాడు** (padi poyAdu) (fell go)
- Conjunct Verb *(Subbarao, 2012)*
  - Light verb that carries tense, aspect, agreement markers, while the semantics is carried by the associated noun/adjective
    - **hi**: mai ne rAma kI madada <u>kI</u>
    - **kn**: nanu ramAnige sahayavannu <u>mAdidene</u>
    - **gloss:** I Ram help did

India as a linguistic area gives us robust reasons for writing a common or core grammar of many of the languages in contact

~ Anvita Abbi

Linguistic Typology

- Genetic Relations
- Contact Relations
- **Linguistic Typology**
- Orthographic Similarity

What is linguistic typology?

- Study of variation in languages & their classification
- Study on the limitations of the degree of variation found in languages

Some typological studies *(Eifring & Theil, 2005)*

- **Word order typology**
- Morphological typology
- Typology of motion verbs
- Phonological typology

Word order typology

- Study of word order in a typical declarative sentence
- Possible word orders:
  - SVO, SOV (85% languages) AND VSO (10% languages)
  - OSV,OVS,VOS (<5% languages)

Correlation between SVO and SOV languages *(Eifring & Theil, 2005)*

<u>SVO Languages</u>

- preposition+noun
  - *<u>in</u> the house*
- noun+genitive or genitive+noun
  - <u>capital</u> of Karnataka
  - Karnataka's capital
- auxilary+verb
  - <u>is</u> coming
- noun+relative clause
  - the <u>cat</u> that ate the rat
- adjective + standard of comparison
  - <u>better</u> than butter

<u>SOV Languages</u>

- noun+postposition
  - घर <u>में</u>
- genitive+noun
  - करनाटक की <u>राजधानी</u>
- verb+auxilary
  - आ <u>रहा है</u>
- relative clause+noun
  - चूहे को खाने वाली <u>बिल्ली</u>
- standard of comparison + adjective
  - मख्खन से <u>बेहतर</u>

In general, it seems head precedes modifier in SVO languages and vice-versa in SOV languages

Orthographic Similarity

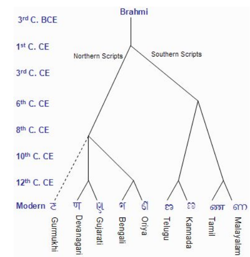
- Genetic Relations
- Contact Relations
- Linguistic Typology
- **Orthographic Similarity**

Writing Systems *(Daniels & Bright, 1995)*

- **Logographic**: symbols representing both sound and meaning
  - Chinese, Japanese Kanji
- **Abjads**: independent letters for consonants, vowels optional
  - Arabic, Hebrew
- **Alphabet**: letters representing both consonants and vowels
  - Roman, Cyrillic, Greek
- **Syllabic**: symbols representing syllables
  - Korean Hangul, Japanese Hiragana & Katakana
- **Abugida**: consonant-vowel sequence as a unit, with vowel as secondary notation
  - Indic Scripts

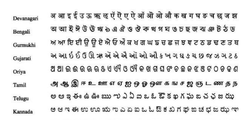
Indic scripts

- All major Indic scripts derived from the *Brahmi* script
  - First seen in Ashoka's edicts
- Same script used for multiple languages
  - Devanagari used for Sanskrit, Hindi, Marathi, Konkani, Nepali, Sindhi, etc.
  - Bangla script used for Assamese too
- Multiple scripts used for same language
  - Sanskrit traditionally written in all regional scripts
  - Punjabi: Gurumukhi & Shahmukhi
  - Sindhi: Devanagari & Persio-Arabic
- Said to be derived from Aramaic script, but shows sufficient innovation to be considered a radically new alphabet design paradigm



Adoption of Brahmi derived scripts

Common characteristics



- *Abugida* scripts: primary consonants with secondary vowels diacritics (*matras*)
  - rarely found outside of the Brahmi family
- The character set is largely overlapping, but the visual rendering differs
- Dependent (maatras) and Independent vowels
- Consonant clusters (क्क,क्ष)
- Special symbols like:
  - *anusvaara* (*nasalization*), *visarga* (aspiration)
  - *halanta/pulli* (vowel suppression), *nukta*(Persian sounds)
- Traditional ordering of characters is same across scripts (*varnamala*)

Organized as per sound phonetic principles

shows various symmetries

Primary vowels

	Short				Long				Diphthongs			
	Initial		Diacritic		Initial		Diacritic		Initial		Diacritic	
Unrounded low central	अ	a	प	pa	आ	ā	पा	pā				
Unrounded high front	इ	i	पि	pi	ई	ī	पी	pī				
Rounded high back	उ	u	पु	pu	ऊ	ū	पू	pū				
Syllabic variants	ऋ	ṛ	पृ	pṛ	ॠ	ṝ	पॄ	pṝ				
	ऌ	ḷ	पॢ	pḷ	ॡ	ḹ	पॣ	pḹ				

Secondary vowels

Unrounded front	ए	e	पे	pe	ऐ	ai	पै	pai
Rounded back	ओ	o	पो	po	औ	au	पौ	pau

Occlusives

	Voiceless plosives				Voiced plosives				Nasals	
	unaspirated		aspirated		unaspirated		aspirated			
Velar	क	ka	ख	kha	ग	ga	घ	gha	ङ	ṅa
Palatal	च	ca	छ	cha	ज	ja	झ	jha	ञ	ña
Retroflex	ट	ṭa	ठ	ṭha	ड	ḍa	ढ	ḍha	ण	ṇa
Dental	त	ta	थ	tha	द	da	ध	dha	न	na
Labial	प	pa	फ	pha	ब	ba	भ	bha	म	ma

Sonorants and fricatives

	Palatal		Retroflex		Dental		Labial	
Sonorants	य	ya	र	ra	ल	la	व	va
Sibilants	श	śa	ष	ṣa	स	sa		

Other letters

ह ha ळ ḷa

Benefits for NLP

- Easy to convert one script to another
- Ensures consistency in pronunciation across a wide range of scripts
- Easy to represent for computation:
  - Coordinated digital representations like Unicode
  - Phonetic feature vectors

Feature	Possible Values
Type	Unused (0), Vowel modifier (1), Nukta (2), Halant (3), Vowel (4), Consonant (5), Number (6), Punctuation (7)
Height (vowels)	Front (1), Mid (2), Back (3)
Length	Short (1), Medium (2), Long (3)
Svar1	Low (1), Lower Middle (2), Upper Middle (3), Lower High (4), High (5)
Svar2	Samvrit (1), Ardh-Samvrit (2) Ardh-Vivrit (3), Vivrit (4)
Sthaan (place)	Dvayoshthya (1), Dantoshthya (2), Dantya (3), Varstya (4), Talavya (5) Murdhanya (6), Komal-Talavya (7), Jivhaa-Muliya (8), Svaryantramukhi (9)
Prayatna (manner)	Sparsha (1), Nasikya (2), Parshvika (3), Prakampi (4), Sangharshi (5), Ardh-Svar (6)

Source: Singh, 2006

- Useful for natural language processing: transliteration, speech recognition, text-to-speech

Some trivia to end this section

The Periodic Table & Indic Scripts

Dmitri Mendeleev is said to have been inspired by the two-dimensional organization of Indic scripts to create the periodic table

http://swarajyamag.com/ideas/sanskrit-and-mendeleevs-periodic-table-of-elements/

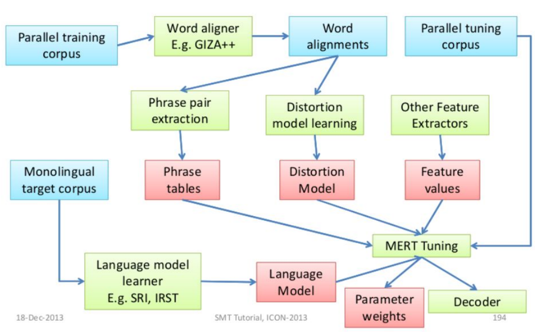
The Full List of Mendeleev's Predictions with their Sanskrit Names

Mendeleev's Given Name	*Modern Name*
Eka-aluminium	Gallium
Eka-boron	Scandium
Eka-silicon	Germanium
Eka-manganese	Technetium
Tri-manganese	Rhenium
Dvi-tellurium	Polonium
Dvi-caesium	Francium
Eka-tantalum	Protactinium

Where are we?

- Motivation
- Language Relatedness
- **A Primer to SMT**
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

The Phrase based SMT pipeline



Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- **Leveraging Orthographic Similarity for transliteration**
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Leveraging Orthographic Similarity for Transliteration

Rule-based transliteration for Indic scripts

(Atreya, et al 2015; Kunchukuttan et al, 2015)

- A naive system: nothing other than Unicode organization of Indic scripts
- First 85 characters in Unicode block for each script aligned
  - Logically equivalent characters have the same offset from the start of the codepage
- Script conversion is simply a question of mapping Unicode characters
- Some exceptions to be handled:
  - Tamil: does not have aspirated and voiceless plosives
  - Sinhala: Unicode codepoints are not completely aligned
  - Some non-standard characters in scripts like Gurumukhi, Odia, Malayalam
- Some divergences
  - Nukta
  - Representation of Nasalization (निशांत or निशान्त)
  - schwa deletion, especially terminal schwa
- This forms a reasonable baseline rule-based system
  - Would work well for Indian origin names
  - English, Persian and Arabic origin have non-standard mappings

Results of Unicode Mapping

	pa	as	bn	hi	gu	mr	te	kn	ml	ta
pa		62.5	87.4	93.2	84.8	66.2	94.3	93.9	94.7	66.2
as	64.8		83.3	72.9	70.5	69.2	64	66.3	60.2	
bn	90.1	82.4		97.3	88.2	64.6	96.1	94.8	98.4	72.9
hi	83.7	71.9	80.9		85.4	76.6	95.9	93.5	95.7	70.7
gu	87.2	71.7	86.6	99		84	97.1	95.4	98	75.2
mr	68.4	71	68	73.2	82.3		64.3	66.8	66.3	
te	97.6	63	97.6	53.8	97	68.2		98.6	99.1	75.1
kn	97.9	64.2	96.1	98.6	96.3	69.7	99.3		99.7	72.2
ml	98.5	61.6	99.3	99.2	98.3	71.4	98.9	99.8		71.4
ta	81.6		81.3	81.7	82		81.1	80.7	79	

Tested on IndoWordNet dataset

Results can be improved can handling the few language specific exceptions that exist

Akshar based transliteration of Indic scripts

(Atreya, et al 2015)

- *Akshar*: A grapheme sequence of the form C+V (क् + त + ई) = क्ती
- An *akshar* approximates a syllable:
  - Syllable: the smallest psychologically real phonological unit (a sound like /kri/)
  - Akshar: the smallest psychologically real orthographic unit (a written akshar like 'kri')
- Vowel segmentation: Segment the word into *akshars*
  - Consider *sanyuktashars* (consonant cluster e.g. *kr*) also as akshars

Hindi	Kannada	English
वि द्या ल य	ವಿ ದಾಯ್ ಲ ಯ	vi dya lay
अ र्जु न	ಅ ರ್ಜು ನ	a rju n

Other possible segmentation methods

Character-based: Split word into characters

Hindi	Kannada	English
व ि द ् य ा ल य	ವ ಿ ದ ್ ಯ ಾ ಲ ಯ	v i d y a l a y
अ र ् ज ु न	ಅ ರ ್ ಜ ು ನ	a r j u n

Syllable-based: Split word at syllable boundaries

- Automatic syllabification is non-trivial
- Syllabification gives best results
- Vowel segmentation is an approximation

Hindi	Kannada	English
विद् या लय	ವಿದ್ ಯಾ ಲ ಯ	vid ya lay
अर् जुन	ಅರ್ ಜು ನ	ar jun

Results for Indian languages

	pa	as	bn	hi	gu	mr	te	kn	ml	ta
pa		CS:77.50 VS:82.50	CS:89.80 VS:93.70	CS:96.80 VS:98.60	CS:90.30 VS:89.50	CS:77.80 VS:78.90	CS:95.70 VS:97.90	CS:96.80 VS:98.40	CS:96.90 VS:98.50	CS:98.50 VS:98.30
as	CS:73.10 VS:83.10		CS:82.58 VS:86.89	CS:76.30 VS:85.90	CS:74.30 VS:84.80	CS:71.00 VS:80.60	CS:71.40 VS:81.70	CS:73.80 VS:85.20	CS:69.00 VS:78.40	-
bn	CS:90.30 VS:93.10	CS:78.60 VS:87.70		CS:97.40 VS:97.80	CS:90.40 VS:93.80	CS:68.20 VS:80.60	CS:96.20 VS:96.90	CS:95.50 VS:97.00	CS:98.40 VS:98.20	CS:97.70 VS:98.00
hi	CS:86.40 VS:87.60	CS:79.30 VS:84.80	CS:79.70 VS:88.30		CS:81.20 VS:88.00	CS:72.77 VS:82.88	CS:95.70 VS:96.50	CS:93.30 VS:93.60	CS:95.40 VS:96.70	CS:95.60 VS:95.80
gu	CS:89.30 VS:88.80	CS:83.00 VS:87.00	CS:84.10 VS:91.20	CS:98.70 VS:99.00		CS:81.60 VS:83.00	CS:97.00 VS:97.00	CS:95.70 VS:96.70	CS:98.00 VS:98.40	CS:98.00 VS:98.20
mr	CS:78.70 VS:79.90	CS:79.40 VS:88.60	CS:75.40 VS:84.40	CS:66.87 VS:75.88	CS:77.40 VS:81.40		CS:67.00 VS:74.60	CS:74.90 VS:78.60	CS:69.20 VS:73.90	-
te	CS:97.40 VS:98.40	CS:75.20 VS:79.80	CS:96.40 VS:98.10	CS:99.20 VS:99.30	CS:97.60 VS:98.20	CS:70.10 VS:76.90		CS:98.70 VS:98.80	CS:99.00 VS:97.70	CS:98.50 VS:98.80
kn	CS:97.60 VS:98.40	CS:76.40 VS:81.30	CS:94.60 VS:97.40	CS:98.50 VS:98.90	CS:96.20 VS:96.80	CS:71.50 VS:79.60	CS:99.20 VS:99.60		CS:99.50 VS:99.90	CS:98.90 VS:99.30
ml	CS:99.00 VS:99.10	CS:72.20 VS:77.70	CS:99.60 VS:99.60	CS:99.10 VS:99.30	CS:98.40 VS:99.00	CS:71.80 VS:77.70	CS:98.90 VS:99.40	CS:99.80 VS:99.90		CS:97.20 VS:97.90
ta	CS:84.10 VS:94.30	-	CS:86.20 VS:95.30	CS:86.80 VS:95.50	CS:86.70 VS:96.60	-	CS:86.50 VS:96.60	CS:86.90 VS:96.20	CS:85.70 VS:95.90	

- Models trained using phrase based SMT system
- Tested on *IndoWordnet* dataset
- **Vowel segmentation outperforms character segmentation**

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- **Leveraging linguistic similarities for translation**
  - **Leveraging Lexical Similarity**
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Leveraging Lexical Similarity

Lexically similar words

Words that are similar in form and meaning

- **Cognates**: words that have a common etymological origin
  - egs. within Indo-Aryan, within Dravidian
- **Loanwords**: borrowed from a donor language and incorporated into a recipient language without translation
  - egs. Dravidian in Indo-Aryan, Indo-Aryan in Dravidian, Munda in Indo-Aryan
- **Fixed Expressions & Idioms**: multiwords with non-compositional semantics
- **Named Entities**

Caveats

- False Friends: words similar in spelling & pronunciation, but different in meaning.
  - Similar origin: semantic shift
  - Different origins pAnl(hi) [water], pani(ml)[fever]
- Loan shifts and other mechanisms of language contact
- Open class words tend to be shared more than closed class words
- Shorter words: difficult to determine relatedness

How can machine translation benefit?

Related languages share vocabulary (cognates, loan words)

- Reduce out-of-vocabulary words & parallel corpus requirements
  - Automatic parallel lexicon (cognates, loan words, named entities) induction
  - Improve word alignment
  - Transliteration is the same as translation for shared words
- Character-oriented SMT

Need a way to measure orthographic & phonetic similarity of words in across languages

Leveraging Lexical Similarity

Reduce OOV words & parallel corpus requirements

- **Phonetic & Orthographic Similarity**
- Identification of cognates & named entities
- Improving word alignment
- Transliterating OOV words

String Similarity Function

If Σ_1 and Σ_2 are alphabet sets and $\Re$ is the real set, a string similarity function can defined as:

$$\text{sim}: \Sigma_1{+} \times \Sigma_2{+} \rightarrow \Re$$

Let's see a few similarity functions

PREFIX *(Inkpen et al,2005)*

- The prefixes of cognates tend to be stable over time
- Compute ratio of matching prefix length to that of longer string

x = "स ◌ु थ ल" **y = "स ◌ु थ ◌ा न"**

***prefix_score(x,y)*=0.6**

- In many cases, the phonetic change in the initial part of the string

x = "अ ◌ं ध ◌ा प न" **y = "आ ◌ं ध ळ ◌े प ण ◌ा"**

***prefix_score(x,y)*=0.0**

Dice & Jaccard Similarity *(Inkpen et al,2005)*

- Bag of word based metrics

  $jaccard(x,y)=|x \cap y| / (|x| + |y| - |x \cap y|)$

  $dice(x,y)= 2*|x \cap y| / (|x| + |y|)$

- Do not take word order into effect

x = "अ ◌ं ध ◌ा प न" **y = "आ ◌ं ध ळ ◌े प ण ◌ा"**

***jaccard(x,y)*=4/10=0.40**

***dice(x,y)* =8/14=0.5714**

LCSR & NED

Metrics that take into account ***order***:

- LCSR: Longest Common Subsequence Ratio *(Melamed, 1995)*

  lcsr(x,y)=ratio of length of longest subsequence to that of longer string
- NED_b: Normalized Edit Distance based metric *(Wagner & Fischer, 1974)*

  ned_b(x,y)=ratio of edit distance to length of longer string

x = “अ ं ध ा प न” **y = “आ ं ध ळ े प ण ा”**

***ned_b(x,y)*=1-(⅝)=0.375**

***lcsr(x,y)*=(3/8)=0.375**

Variants

- Instead of unigrams, n-grams could be considered as basic units. Favours matched characters to be contiguous *(Inkpen et al,2005)*

  x = “अ ं ध ा प न” **y = “आ ं ध ळ े प ण ा”**

  dice_2gram(x,y) **=1/12=8.33**

- *Skip gram based metrics* could be defined by introducing gaps *(Inkpen, 2005)*
- Use *similarity matrix* to encode character similarity, substitution cost
- Learn similarity matrices automatically *(Ristad, 1999; Yarowsky, 2001)*
- LCSF metric to fix LCSR preference for short words *(Kondrak, 2005)*

Phonetic Similarity & Alignment

Given a pair of phoneme sequences, find the alignment between the phonemes of the two sequences, and an alignment score:

अ न् ध ा - - प न - (andhApana, Hindi)

आ न् ध - ळ े प ण ा (AndhaLepaNA, Marathi)

assuming the Indic script characters to be equivalent to phonenems, else represent the examples using IPA

You need the following:

- Grapheme sequence to phoneme sequence conversion
- Mapping of phonemes to their phonetic features
- Phoneme Similarity function
- Algorithm for computing alignment between the phoneme sequence

Phonetic Feature Representation for phonemes

Feature	Values
Basic Character Type	vowel , consonant, nukta, halanta, anusvaara, miscellaneous
Vowel Length	short, long
Vowel Strength	weak (a,aa,i,ii,u,uu), medium (e,o), strong (ai,au)
Vowel Status	Independent, Dependent
Consonant Type	plosive (क to म), fricative (स,ष,श,ह), central approximant(य,व,zha), lateral approximant (la,La), flap(ra,Ra)
Place of Articulation	velar, palatal, retroflex, dental, labial
Aspiration	True, False
Voicing	True, False
Nasal	True, False

Phonetic Similarity Function

If **P** is set of phonemes and $\Re$ is the real set, a similarity function is defined as:

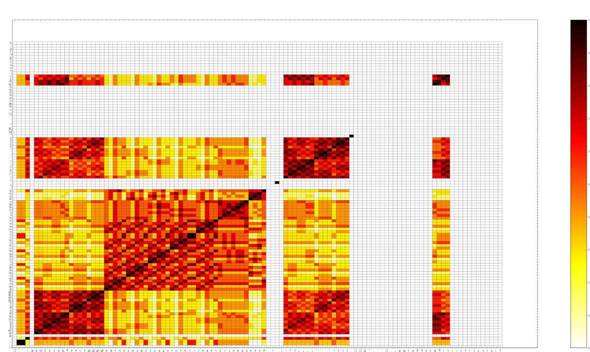
sim: $\mathbf{P}\times\mathbf{P} \rightarrow \Re$

Or a corresponding distance measure could be defined

Some common similarity functions

- Cosine similarity
- Hamming distance
- Hand-crafted similarity matrices

Cosine similarity



Phonemic similarity between Devanagari characters

Multi-valued features and similarity

Some feature values are similar to each other than others

- Labio-dental sounds are more similar to bilabial sounds than velar sounds
- Weights are assigned to each possible value a feature can take
- **Difference in weights can capture this intuition**

Feature name	Phonological term	Numerical value
Place	[bilabial]	1.0
	[labiodental]	0.95
	[dental]	0.9
	[alveolar]	0.85
	[retroflex]	0.8
	[palato-alveolar]	0.75
	[palatal]	0.7
	[velar]	0.6
	[uvular]	0.5
	[pharyngeal]	0.3
	[glottal]	0.1
Manner	[stop]	1.0
	[affricate]	0.9
	[fricative]	0.8
	[approximant]	0.6
	[high vowel]	0.4
	[mid vowel]	0.2
	[low vowel]	0.0
High	[high]	1.0
	[mid]	0.5
	[low]	0.0
Back	[front]	1.0
	[central]	0.5
	[back]	0.0

Source: Kondrak, 2000

Some features are more important than others

Covington's distance measure
Covington (1996)

	Clause in Covington's distance function	Covington's penalty
1	*"identical consonants or glides"*	0
2	*"identical vowels"*	5
3	*"vowel length difference only"*	10
4	*"non-identical vowels"*	30
5	*"non-identical consonants"*	60
6	*"no similarity"*	100

Source: Kondrak, 2000

Features used in in ALINE & salience values
Kondrak (2000)

Syllabic	5	Place	40
Voice	10	Nasal	10
Lateral	10	Aspirated	5
High	5	Back	5
Manner	50	Retroflex	10
Long	1	Round	5

Source: Kondrak, 2000

Alignment Algorithm

- Standard Dynamic-Programming algorithm for local alignment like Smith-Waterman
- Can extend it to allow for expansions, compressions, gap penalties, top-n alignments
- The ALINE algorithm *(Kondrak, 2000)* incorporates many of these ideas

A matrix H is built as follows:

$$H(i,0) = 0,\ 0 \le i \le m$$

$$H(0,j) = 0,\ 0 \le j \le n$$

$$H(i,j) = \max \left\{ \begin{array}{lc} 0 & \\ H(i-1,j-1) + s(a_i,b_j) & \text{Match/Mismatch} \\ \max_{k \ge 1}\{H(i-k,j) + W_k\} & \text{Deletion} \\ \max_{l \ge 1}\{H(i,j-l) + W_l\} & \text{Insertion} \end{array} \right\},\ 1 \le i \le m, 1 \le j \le n$$

Where:

- a, b = Strings over the Alphabet Σ
- $m = \text{length}(a)$
- $n = \text{length}(b)$
- $s(a,b)$ is a similarity function on the alphabet
- $H(i,j)$ - is the maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j]
- W_i is the gap-scoring scheme

Source: Wikipedia

Leveraging Lexical Similarity

Reduce OOV words & parallel corpus requirements

- Phonetic & Orthographic Similarity
- **Identification of cognates & named entities**
- Improving word alignment
- Transliterating OOV words

Methods

Thresholding based on similarity metrics

Classification with similarity & other features

Competitive Linking

Features for a Classification System

- String (*LCSR, NED_b, PREFIX, Dice, Jaccard, etc.)* & Phonetic Similarity measures *(Bergsma & Kondrak, 2007)*
- Aligned n-gram features *(Klementiev & Roth, 2006; Bergsma & Kondrak, 2007)*

  (पानी,पाणी) → (प,प),(◌ा,◌ा),(◌ी,◌ी)   (पा,पा)

- Unaligned n-gram features *(Bergsma & Kondrak, 2007)*

  (पानी,पाणी) → (न,ण),(◌ानी,◌ाणी)

- Contextual similarity features

Competitive Linking (Melamed, 2000)

- Meta-algorithm which can be used when pairwise scores are available
- Represent candidate pairs by a complete bipartite graph
  - Edge weights represents score of the candidate cognate pairs
- Solution: Find maximum weighted matching in the bipartite graph
- NP-complete
- Heuristic solution:
  - Find candidate pair with maximum association
  - Remove these from further consideration
  - Iterate

Cognates/False-friends vs. Unrelated (Inkpen *et al* 2005)

Orthographic similarity measure	Threshold	Accuracy
IDENT	1	43.90%
PREFIX	0.03845	92.70%
DICE	0.29669	89.40%
LCSR	0.45800	92.91%
NED	0.34845	93.39%
SOUNDEX	0.62500	85.28%
TRI	0.0476	88.30%
XDICE	0.21825	92.84%
XXDICE	0.12915	91.74%
BI-SIM	0.37980	94.84%
BI-DIST	0.34165	94.84%
TRI-SIM	0.34845	95.66%
TRI-DIST	0.34845	95.11%
Average measure	0.14770	93.83%

Performance of individual measures
Thresholds were learnt using single feature classifier

Classifier	Accuracy cross-val
Baseline	63.75%
OneRule	95.66%
Naive Bayes	94.84%
Decision Trees	95.66%
DecTree (pruned)	95.66%
IBK	93.81%
Ada Boost	95.66%
Perceptron	95.11%
SVM (SMO)	95.46%

Results of classification

- LCSR, NED are simple, effective measures
- n-gram measures perform well
- Classification gives modest improvement over individual measures on this simple task

Cognate vs False Friend (Bergsma & Kondrak (2007))

Individual measures

Learning Similarity

Classification

	Bitext			Dictionary					
System	Fr	Es	De	Fr	Es	De	Gr	Jp	Rs
PREFIX	34.7	27.3	36.3	45.5	34.7	25.5	28.5	16.1	29.8
DICE	33.7	28.2	33.5	44.3	33.7	21.3	30.6	20.1	33.6
LCSR	34.0	28.7	28.5	48.3	36.5	18.4	30.2	24.2	36.6
NED	36.5	**31.9**	32.3	50.1	**40.3**	23.3	**33.9**	28.2	41.4
PREFIX+DICE+LCSR+NED	**38.7**	31.8	**39.3**	**51.6**	40.1	**28.6**	33.7	22.9	37.9
Kondrak (2005): LCSF	29.8	28.9	29.1	39.9	36.6	25.0	30.5	**33.4**	**45.5**
Ristad & Yanilos (1998)	37.7	32.5	34.6	56.1	46.9	36.9	38.0	52.7	51.8
Tiedemann (1999)	38.8	33.0	34.7	55.3	49.0	24.9	37.6	33.9	45.8
Klementiev & Roth (2006)	61.1	55.5	53.2	73.4	62.3	48.3	51.4	62.0	64.4
Alignment-Based Discriminative	**66.5**	**63.2**	**64.1**	**77.7**	**72.1**	**65.6**	**65.7**	**82.0**	**76.9**

Bitext, Dictionary Foreign-to-English cognate identification 11-pt average precision (%).

- More difficult task
- LCSR, NED are amongst the best measures
- Learning similarity matrices improves performance
- Classification based methods outperform other methods

Leveraging Lexical Similarity

Reduce OOV words & parallel corpus requirements

- Phonetic & Orthographic Similarity
- Identification of cognates & named entities
- **Improving word alignment**
- Transliterating OOV words

Augmenting Parallel Corpus with Cognates

Add cognate pairs to the parallel corpus

Heuristics

- High recall cognate extraction better than high precision *(Kondrak et al, 2003; Onaizan, 1999)*
  - alignment methods robust to some false positive among cognate pairs
- Replication of cognate pairs improves alignment quality marginally *(Kondrak et al, 2003; Och & Ney, 1999; Brown et al, 1993)*
  - Higher replication factors for words in training corpus to avoid topic drift
  - Replication factor can be elegantly incorporated into the word alignment models
- One vs multiple cognate pairs per line
  - better alignment links between respective cognates for multiple pairs per line *(Kondrak et al, 2003)*

Augmenting Parallel Corpus with Cognates (2)

Results from *Kondrak et al (2003)*

- *Implicitly improves word alignment*: 10% reduction of the word alignment error rate, from 17.6% to 15.8%
- *Improves vocabulary coverage*
- *Improves translation quality*: 2% improvement in BLEU score

Evaluation	Baseline	Cognates
Completely correct	16	21
Syntactically correct	8	7
Semantically correct	14	12
Wrong	62	60
Total	100	100

- Cannot translate words not in parallel or cognate corpus
- Knowledge locked in cognate corpus is underutilized

This method is just marginally useful

Using orthographic features for Word Alignment

- Generative IBM alignment models can't incorporate phonetic information
- Discriminative models allow incorporation of arbitrary features *(Moore, 2005)*
- Orthographic features for English-French word alignment: *(Taskar et al, 2005)*
  - exact match of words
  - exact match ignoring accents
  - exact matching ignoring vowels
  - LCSR
  - short/long word
- **7% reduction in alignment error rate**
- Similar features can be designed for other writing systems
- Cannot handle OOVs

Model	AER
Dice (without matching)	38.7 / 36.0
Model 4 (E-F, F-E, intersected)	8.9 / 9.7/ 6.9
Discriminative Matching	
Dice Feature Only	29.8
+ Distance Features	15.5
+ Word Shape and Frequency	14.4
+ Common Words and Next-Dice	10.7
+ Model 4 Predictions	5.4

Word Error Rates of English-French word alignment task (Taskar et al, 2005)

Leveraging Lexical Similarity

Reduce OOV words & parallel corpus requirements

- Phonetic & Orthographic Similarity
- Identification of cognates & named entities
- Improving word alignment
- **Transliterating OOV words**

Transliterating OOV words

- OOV words can be:
  - **Cognates**
  - **Loan words**
  - **Named entities**
  - Other words
- Cognates, loanwords and named entities are related orthographically
- *Transliteration achieves translation*
- Orthographic mappings can be learnt from a parallel transliteration/cognate corpus

Transliteration as Post-translation step

Durrani et al (2014), Kunchukuttan et al (2015)

Option 1: Replace OOVs in the output with their best transliteration

Option 2: Generate top-k candidates for each OOV. Each regenerated candidates is scored using an LM and the original features

Option 3: 2-pass decoding, where OOV are replaced by their transliterations in second pass input

Rescoring with LM & second pass use LM context to disambiguate among possible transliterations

Translate vs Transliterate conundrum

False friends hi: mujhe pAnI cahiye (I want water) ml-xlit-OOV : enikk paNi vennum (I want work) ml: enikk veLL.m vennum	Name vs word en: Bhola has come home hi: bholA ghara AyA hai en: The innocent boy has come home hi: vah bholA ladkA ghara AyA hai
Which part of a name to transliterate? United Arab Emirates s.myukta araba amirAta	Transliteration is not used United States amrIkA

Integrate Transliteration into the Decoder

Durrani et al (2010), Durrani et al (2014)

- In addition to translation candidates, decoder considers all transliteration candidates for each word
  - Assumption: 1-1 correspondence between words in the two languages
  - monotonic decoding
- Translation and Transliteration candidates compete with each other
- The features used by the decoder (LM score, factors, etc.) help make a choice between translation and transliteration, as well as multiple transliteration options

Additional Heuristics

1. **Preferential treatment for true cognates**: Reinforce cognates which have the same meaning as well as are orthographically similar using new feature:

   *joint_score(f,e) = sqrt(xlation_score(f,e) * xlit_score(f,e))*

2. **LM-OOV feature**:
   - Number of words unknown to LM.
   - Why?: LM smoothing methods assign significant probability mass to unseen events
   - This feature penalizes such events

Results (Hindi-Urdu Translation)

Durrani et al (2010)

Phrase-Based (1)	(1)+Post-edit Xlit	(1)+PB with in-decoder Xlit (3)	(3) + Heuristic 1
14.3	16.25	18.6	18.86

Hindi and Urdu are essentially literary registers of the same language. We can see a 31% increase in BLEU score

फिर भी वह शान्ती से नहीं रह सकता है
پھر بھی وہ سکون سے نہیں رہ سکتا ہے
p_hIr b_hi vh s@kun se n@herh s@kt_dA
"Even then he can't live peacefully"
ओम शान्ती ओम फराह खान की दूसरी फिल्म है
اوم شانتی اوم فراح خان کی دوسری فلم ہے
Aom SAnt_di Aom frhA xAn ki d_dusri fIl@m he
"Om Shanti Om is Farah Khan's second film"

Transliteration Post-Editing for Indian languages

Kunchukuttan et al (2015)

	hi	ur	pa	bn	gu	mr	kK	ta	te	ml	en
hi	-	19.26	23.98	21.05	21.25	19.87	18.39	9.84	15.38	11.47	8.25
ur	16.67	-	17.65	26.32	10.53	9.52	11.11	13.04	14.29	4.35	5.56
pa	29.54	20.14	-	20.62	20.53	17.40	16.90	6.87	14.18	7.55	6.55
bn	27.35	17.17	22.57	-	22.01	20.05	19.19	7.68	14.96	10.38	8.41
gu	33.82	21.67	27.34	25.72	-	25.82	22.15	8.66	17.66	10.54	7.68
mr	30.29	17.50	23.77	25.08	29.07	-	25.25	8.79	16.50	9.54	4.99
kK	27.89	18.21	23.81	23.96	24.01	24.21	-	9.29	16.17	10.17	6.05
ta	16.90	11.38	12.40	13.63	13.07	11.00	11.82	-	11.32	8.67	3.64
te	19.53	11.49	16.74	15.59	15.00	13.20	13.02	7.36	-	7.73	5.07
ml	15.50	8.95	11.70	13.22	12.26	10.14	10.39	7.94	10.97	-	3.54
en	5.85	5.22	4.70	4.16	3.34	3.11	4.34	1.91	4.11	2.79	-

% decrease in OOV using statistical transliteration

- Transliterate untranslated words & rescore with LM and LM-OOV features (Durrani, 2014)
- BLEU scores improve by up to 4%
- OOV count reduced by up to 30% for IA languages, 10% for Dravidian languages
- Nearly correct transliterations: another 9-10% decrease in OOV count can potentially be obtained

Leveraging Lexical Similarity

Character-oriented SMT (CO-SMT)

Key ideas

- **Translation as Transliteration**
- Character as the basic unit of translation
- Represent the sentence as a pair of character sequence
- Word boundaries are represented by special characters

Example

word-level representation
(hi) राम ने श्याम को पुस्तक दी
(mr) रामाने श्यामला पुस्तक दिली

char-level representation
(hi) र ा म _ न े _ श ् य ा म _ क ो _ प ु स ् त क _ द ी
(mr) र ा म ा न े _ श ् य ा म ल ा _ प ु स ् त क _ द ि ल ी

Motivation (Neubig et al, 2012)

- The primary divergences between related languages/dialects are:
  - spelling/pronunciation differences
  - suffix sets
  - function words
- A single integrated framework to tackle:
  - Named entities
  - Cognates
  - High degree of inflection and agglutination
  - Lack of word boundaries
- In short, handle data sparsity is the issue
- *Can this concept apply to any pair of languages?*

Making CO-SMT work

Corpus representation: Add word-boundary boundary marker character

Sentences are too long; decoding and word alignment are inefficient

- Limit on sentence length in training corpus; loss of training corpus *(Tiedemann, 2009)*
- Extract phrases from word based phrase table as candidates; larger models *(Vilar, 2007)*

No distinct advantage of one model over another *(Tiedemann, 2009)*

Limitations:

- Does not solve the decoding problem
- Is the corpus representative?

Monotone decoding: since character level reordering is not properly defined. However, using reordering has also been shown to be useful *(Tiedemann, 2009)*

Tuning: character level tuning not meaningful, should be done at the word level *(Tiedemann, 2012)*

Squeezing out performance from CO-SMT

Capturing larger context information *(Tiedemann, 2009)*

- Larger order LM
- Larger phrase lengths

Viable since data sparsity is not an issue in the character space (except for logographic scripts). Improves translation accuracy.

Exploring the character → word oriented translation continuum

Overlapping n-gram as basic unit *(Tiedemann, 2012)*

Combining with a word-oriented SMT (WO-SMT) *(Nakov & Tiedemann, 2012)*

- System combination of CO-SMT and WO-SMT and selecting translation outputs
- Merging the two models:
    - transform WO-SMT phrase table to character level
    - Add origin features

Results

System	BLEU%	LCSR%
word-based (lexicalised reord)	**50.12**	75.95
char-based (lexicalised reord)	48.98	80.65
char-based (monotone)	48.94	80.36
char-based (lexicalised reorder) +longer n-gram & phrase length	50.07	**80.94**

Source: Tiedemann, 2009
Norwegian→ Swedish translation

No	System	%BLEU
1	word-based	32.19
2	char-based (unigram)	32.28
3	char-based (bigram)	32.71
4	system combination (MEMT) (3+4)	32.92
5	merging phrase tables (4+4)	33.94

Source: Nakov & Tiedemann, 2012 for
Macedonian→ Bulgarian translation

- As measured by BLEU metric, character based models are comparable word level models
    - BLEU is not an appropriate metric, since exact words may not be generated
    - Evaluator can still perceive good translation quality, LCSR may capture that better
- Longer LM and phrase context in char based model helps
- Combining word based and character based models improves translation accuracy

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- **Leveraging linguistic similarities for translation**
    - Leveraging Lexical Similarity
    - **Leveraging Morphological Similarity**
    - Leveraging Syntactic Similarity
- Synergy among multiple languages
    - Pivot-based SMT
    - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Morphological Similarities

Word segmentation improves translation output for morphologically rich languages

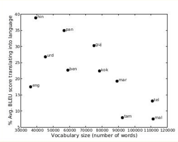
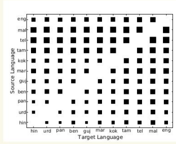
Morphological Similarity

- Related languages may exhibit **morphological isomorphism**
    - correspondence between the suffixes and post-positions
    - e.g. source suffix → target suffix + target post-position
      വീടിനു മുന്നിൽ (vITinu munnil)→ घर के सामने (ghar ke sAmne) (in front of the house)
- Isomorphism makes translation easier
    - If suffixes were translated as phrases, these would have to be learnt from parallel corpus
- Morphological divergences to be bridged
    - Does the source suffix transform to target suffix or post-position or both?
    - Are there multiple options for translation of the suffix?

The challenge of morphological complexity

- Too many unique words
- Translation probabilities cannot be learnt reliably
- Many words are not translated; OOVs in translation output

(Kunchukuttan et al 2014 (a))

Increased Morphological complexity decreases translation accuracy:
- Strong inverse correlation between corpus vocabulary size and average BLEU score translating into a language ($\mathbf{r} = -\mathbf{0.7}$)
- Marathi & Konkani: Lower BLEU scores for the morphologically richest Indo-Aryan languages
- Translation Model Entropy (TME): Uncertainty in selecting a translation of a source phrase
  - High TME for SMT systems involving morphologically rich languages
  - Low TME for Indo-Aryan, high TME for Dravidian language pairs

Unsupervised Word Segmentation

Reduce data sparsity by decomposing words in training corpus into their component morphemes

മംഗൾയാൻ ഒമ്പത് മാസങ്ങൾകഴിഞ്ഞ് ചൊവ്വയിൽ എത്തി
maMgaLyAn ompata mAsa.NgaL kazhiJN chovvayil etti
Mangalyan nine months after Mars_in reached

മംഗൾയാൻ ഒമ്പത് മാസ ങ്ങൾ കഴിഞ്ഞ് ചൊവ്വ യിൽ എത്തി
maMgaLyAn ompata mAsa .NgaL kazhiJN chovva yil etti

- Learn word segmentation from a list of words and their corpus frequencies (optional)
- Finds the lexicon (set of morphemes) such that the following objectives are met:
  - The likelihood of the tokens is maximized
  - The size of lexicon is minimized
  - Shorter morphemes are preferred
- The technique is language independent and requires and only monolingual resources to learn word segmentation

മംഗൾയാൻ ഒമ്പത് മാസങ്ങൾ കഴിഞ്ഞ് ചൊവ്വയിൽ എത്തി
maMgaLyAn ompata mAsa.NgaL kazhiJN chovvayil etti
Mangalyan nine months after Mars_in reached

Morphological Segmentation

മംഗൾയാൻ ഒമ്പത് മാസ ങ്ങൾ കഴിഞ്ഞ് ചൊവ്വ യിൽ എത്തി
maMgaLyAn ompata mAsa .NgaL kazhiJN chovva yil etti

Translate morph-segmented Malayalam to Hindi

മംഗൾയാൻ नौ महीने बाद मंगल पहुँचा
maMgaLyAn nau mahIne bAd mangal pah.Ncha

Generate transliteration candidates for untranslated words

मंगलयान/मंगालयान/मँगलयान नौ महीने बाद मंगल पहुँचा
maMgalyAn/maMgAlyAn/ma.NgalyAn nau mahIne bAd mangal pah.Ncha

Select best candidate sentence

मंगलयान नौ महीने बाद मंगल पहुँचा
maMgalyAn nau mahIne bAd mangal pah.Ncha
Mangalyan nine months after Mars reached

- Word segmentation makes it possible to align segments from the language pairs involved
- Because of similarity of morphological properties, correspondences between morphemes on either side can be easily found

Results for IL-hi translation (Kunchukuttan et al 2014 (b))

Lang Pair	Metric	Tourism PB	Tourism PB+ morph	Tourism PB+ morph+ translit	Health PB	Health PB+ morph	Health PB+ morph+ translit	General PB	General PB+ morph	General PB+ morph+ translit
bn-hi	B	34.38	37.1	**37.66**	36.46	38.66	**39.04**	36.24	38.61	**38.92**
	M	55.73	58.38	**58.98**	57.44	59.89	**60.37**	57.36	59.47	**59.84**
mr-hi	B	40.24	**46.86**	**46.86**	39.84	46.86	**46.86**	41.35	47.92	**47.92**
	M	60.78	**66.47**	**66.47**	60.29	**66.76**	**66.76**	61.79	**67.17**	**67.17**
ta-hi	B	17.76	22.42	**22.91**	21.55	26.05	**26.35**	20.45	25.34	**25.65**
	M	36.11	41.61	**42.31**	39.94	45.03	**45.42**	38.93	44.57	**50.00**
te-hi	B	26.99	31.77	**32.45**	29.74	35.59	**36.04**	29.88	35.43	**35.88**
	M	47.20	52.48	**53.35**	50.05	56.05	**56.68**	50.20	55.82	**56.38**

- Source word segmentation significantly improves performance
  - For morphologically rich source like *ta*, improvements of upto 24% in BLEU
  - For comparatively poor source like *bn*, improvements of upto 6% in BLEU
  - Similar trends for METEOR score
- Transliteration post-editing marginally improves translation
  - BLEU scores improve by upto 1.2%
  - Recall improves by upto 1.4%

Examples

Morphological segmentation helps overcome data sparsity

Source	गौतम बुद्ध अभयारण्य कोडरमामध्ये वसलेले आहे जेथे चित्ता आणि वाघ आहेत .
Segmented	गौतम बुद्ध अभयारण्य कोडरमा मध्ये वसलेल े आहे जेथे चित्ता आणि वाघ आहेत .
Xlation: simple PBSMT	गौतम बुद्ध अभ्यारण्य कोडरमामध्ये स्थित है जहाँ चीता और बाघ हैं ।
Xlation: PBSMT + segmentation	गौतम बुद्ध अभ्यारण्य कोडरमा में स्थित है जहाँ चीता और बाघ हैं ।

Aggressive segmentation results in deterioration of translation quality

Source	इक्ष्वाकु पुत्र राजा विशाल याला वैशाली राज्याचा संस्थापक मानले जाते .
Segmented	इ क्ष ्वा कु पुत्र राजा विशाल याला वैशाली राज्य ाचा संस्थापक मानले जाते .
Xlation: simple PBSMT	इक्ष्वाकु पुत्र राजा विशाल इसे वैशाली राज्य का संस्थापक माना जाता है ।
Xlation: PBSMT + segmentation	सन सफेद ्वा विकृत पुत्र राजा विशाल इसे वैशाली राज्य का संस्थापक माना जाता है ।

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- **Leveraging linguistic similarities for translation**
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - **Leveraging Syntactic Similarity**
- Synergy among multiple languages
  - Pivot-based SMT
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Syntactic Similarities

Source reordering for English → Indian language SMT

The structural divergence problem for En-IL

- Significant structural divergence between English and Indian languages (Indo-Aryan & Dravidian)
  - English is SVO
  - All Indian languages are SOV
- Standard PBSMT cannot handle long-distance reordering
- *Source Reordering*: Change the word of source side of the training corpus to match the target language word order prior to SMT training

English	$\overbrace{\textit{The president}}^{S}\ \overbrace{\textit{of America}}^{S_m}\ \overbrace{\textit{visited}}^{V}\ \overbrace{\textit{India}}^{O}\ \overbrace{\textit{in June}}^{V_m}$
Reordered	$\underbrace{\textit{America of}}_{S_m}\ \underbrace{\textit{the president}}_{S}\ \underbrace{\textit{June in}}_{V_m}\ \underbrace{\textit{India}}_{O}\ \underbrace{\textit{visited}}_{V}$
Hindi	अमरीका के राष्ट्रपति ने जून में भारत की यात्रा की amariikaa ke raashtrapati ne juuna mem bhaarata kii yaatraa kii

- Source Reordering improves PBSMT:
  - Longer phrases can be learnt
  - Decoder cannot evaluate long distance reorderings by search in a small window

Rule-based source reordering

Generic reordering *(Ramanathan et al 2008)*

Basic reordering transformation for English→

$$SS_mVV_mOO_mCm \rightarrow C'_mS'_mS'O'_mO'V'_mV'$$

where,
S: Subject
O: Object
V: Verb
C_m: Clause modifier
X': Corresponding constituent in Hindi, where X is S, O, or V
X_m: modifier of X

Hindi-tuned reordering *(Patel et al 2013)*

Improvement over the basic rules by analyzing En→ Hi translation output

VP(advP vpw dcP: advP dcP vpw)

English: Bikaner, popularly **known as the camel county** is located in Rajasthan.

Parse: Bikaner , *[VP (advP* popularly*) (vpw* known*) (dcP* as the camel country*)]* is located in Rajsthan .

Partial Reordered: Bikaner , *(advP* popularly*)* ***(dcP* as the camel country*) (vpw* known*)*** is located in Rajsthan .

Reordered: Bikaner , *(advP* popularly*) (dcP* the camel country as*) (vpw* known*)* Rajsthan in located is .

Hindi: *bikaner , jo aam taur par unton ke desh ke naam se jana jata hai, rajasthan me sthit hai .*

Portable rules for En⟶ IL pairs

	Indo-Aryan							Dravidian			
	hin	urd	pan	ben	guj	mar	kok	tam	tel	mal	eng
(A) Phrase based system (S1)											
eng	26.53	18.07	22.86	14.85	17.36	10.17	13.01	4.17	6.43	4.85	-
(B) Phrase based system with source reordering: generic rules (S2)											
eng	29.63	20.42	26.06	16.85	20.11	11.46	15.01	4.97	7.83	5.53	-
(C) Phrase based system with source reordering: Hindi-adapted rules (S3)											
eng	30.86	21.54	27.52	18.20	21.33	12.68	15.73	5.09	8.29	5.68	-

S2: Generic En-Hi reordering rule-base

S3: En-Hi reordering rule-base, tuned for Hindi

- Source reordering improves BLEU scores for 15% and 21% for source reordering system systems S2 and S3 respectively for all language pairs
- **A single rule-base serves all major Indian languages**
- Even Hindi-tuned rules perform well for other Indian languages as target

Examples

Source reordering helps improves word order

Steps	Sentence
Input Sentence	Bilirubin named colored substance is made in our body absolutely everyday .
Source side reordering	Bilirubin named colored substance in our body absolutely everyday made is .
Phrase based Translation	Bilirubin नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।
Transliteration	वाइलीरुविन नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।

Reordering rules can generate wrong word order

In this example, no rules for imperative sentences cause reordering error

Input Sentence	Burn on cooking 20 live scorpions in 1 litre sesame seed oil .
Source side reordering	1 in 20 live scorpions cooking on Burn sesame seed oil litre .

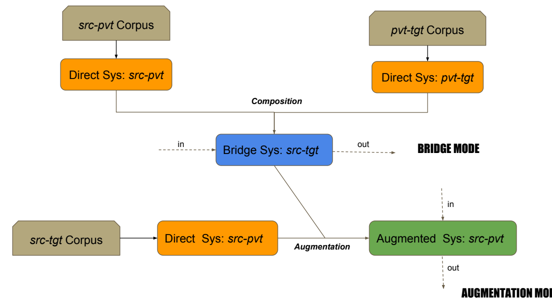
Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- **Synergy among multiple languages**
  - **Pivot based SMT**
  - Multi-source translation
- Summary & Conclusion
- Tools & Resources

Pivot based SMT

- **Core concepts**
- What is a good pivot?
- Addressing language divergences in pivot based SMT

Translation using pivot languages



Why pivot based SMT?

Bridge Mode

No parallel resources are available between source and target languages

Augmentation Mode

Scarce parallel resources between source and target languages, but ample resources between source-pivot and/or pivot/target

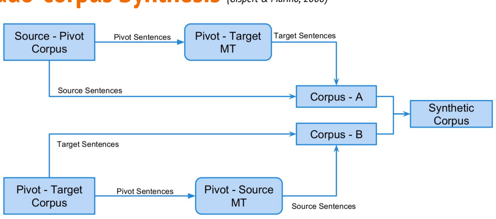
- New translation pairs
- New translation options

Improvement in lexical coverage

Methods for Composition of *src-pvt* and *pvt-tgt* systems

- Pseudo-Corpus Synthesis
- Cascading Direct Systems
- Model Triangulation

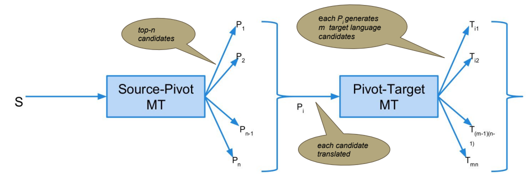
Pseudo-corpus Synthesis *(Gispert & Marino, 2006)*



Source: More, 2015

- Either Corpus A or Corpus B can be used or both can be used
- Generated corpus will be noisy: quality would depend on the divergence between the language pairs and the size of the parallel corpus
- Easy to implement
- Same runtime complexity as a single model

Cascading Direct Systems *(Utiyama & Isahara, 2007)*



Source: More, 2015

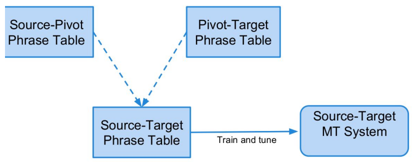
- Rank the *m.n* target language candidates using:

$$\hat{t} = \underset{t}{\operatorname{argmax}} \sum_{k=1}^{L} \left(\lambda_k^{sp} h_k^{sp}(s,p) + \lambda_k^{pt} h_k^{pt}(p,t)\right)$$

where, (i) L is number of features, (ii) λ's are feature weights, (iii) h's are feature values (iv) *sp, pt:* src-pvt & pvt-tgt models

- Easy to implement
- Compute intensive: n+1 decoding runs per sentence
- top-n configuration is generally better than top-1

Model Triangulation *(Utiyama & Isahara, 2007; Wu & Wang, 2007)*



Source: More, 2015

- Merges the Source-Pivot and Pivot-Target models
- In a phrase based settings, this means:
  - Merge Phrase Tables and induce feature values (phrase translation & lexical probability)
  - Merge Reordering Tables
- The merge can be motivated in a systematic & elegant manner from elementary probability theory
- The size of the resultant tables is much larger than input tables
- The best performing method

Model Triangulation Explained

Given: Source-Pivot and Pivot-Target Phrase tables

Goal: Merge the two into a single phrase table, and compute the feature values:

- Phrase translation probability
- Lexical probability

Like performing a database join, but the feature values also have to be merged

src-pivot table

A	X	0.1	0.4
B	X	0.6	0.8
B	Y	0.8	0.9
C	Y	0.3	0.4

pivot-tgt table

X	P	0.5	0.4
Y	P	0.9	0.7
Y	Q	0.1	0.9
Z	R	0.3	0.7

A	P	?	?
B	P	?	?
B	Q	?	?
C	Q	?	?
C	P	?	?

Table based approach for computing probabilities

Utiyama & Isahara, 2007

src-pivot table

A	X	0.1	0.4
B	X	0.6	0.8
B	Y	0.8	0.9
C	Y	0.3	0.4

pivot-tgt table

X	P	0.5	0.4
Y	P	0.9	0.7
Y	Q	0.1	0.9
Z	R	0.3	0.7

A	P	0.05	0.16
B	P	0.51	0.475
B	Q	0.08	0.81
C	Q	0.03	0.36
C	P	0.27	0.28

To computing phrase & lexical translation probability, marginalize over all pivots phrases

$$\phi(\bar{s}|\bar{t}) = \sum_{\bar{p}} \phi(\bar{s}|\bar{t},\bar{p})\phi(\bar{p}|\bar{t}) \qquad p_w(\bar{s}|\bar{t}) = \sum_{\bar{p}} p_w(\bar{s}|\bar{p},\bar{t})p_w(\bar{p}|\bar{t})$$

Since the source phrase is independent of the target given the pivot,

$$\phi(\bar{s}|\bar{t}) = \sum_{\bar{p}} \phi(\bar{s}|\bar{p})\phi(\bar{p}|\bar{t}) \qquad p_w(\bar{s}|\bar{t}) = \sum_{\bar{p}} p_w(\bar{s}|\bar{p})p_w(\bar{p}|\bar{t})$$

s, t, p are source, target and pivot, phrases respectively
ϕ: phrase translation probability
p_w: lexical translation probability

The terms on the right can be obtained from src-pvt and pvt-tgt phrase tables respectively

Count based method for lexical probability

(Wu & Wang, 2007)

Lexical probability is computed from words alignments as:

$$p_w(\bar{s}|\bar{t},a) = \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{\forall (i,j)\in a} w(s_i|t_j)$$

Induce source-target alignments from alignments in the original phrase tables

src-pivot table

	s1	s2	s3
p1	*		
p2		*	
p3			*

pivot-tgt table

	p1	p2	p3
t1	*		
t2		*	
t3		*	

	s1	s2	s3
NULL			*
t1	*		
t2		*	
t3		*	

Count based method for lexical probability (2)

Now count the co-occurrence of (src,pvt) words in induced alignments

$$count(s,t) = \sum_{k=1}^{K} \phi_k(\bar{s}|\bar{t}) \sum_{i=1}^{n_k} \delta(s,s_i)\delta(t,t_{a_i})$$

The counts in each phrase are weighted by the phrase translation probability

Now compute the word translation probability

$$w(s|t) = \frac{\text{count}(s,t)}{\sum_{s'} \text{count}(s',t)}$$

Another method to compute *w* (Wang, 2006), where *sim* is cross language word similarity

$$w(s|t) = \sum_{p} w(s|p)w(p|t)sim(s,t;p)$$

Now plug these values back into equation for lexical probability

Count based better than similarity based

$$p_w(\bar{s}|\bar{t},a) = \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{\forall (i,j)\in a} w(s_i|t_j)$$

Comparison of Composition Methods

Criteria	Pseudo-corpus	Cascaded	Triangulation
Ease of implementation	Easy	Easy	Involved
Training Time	Low, just as much as a baseline PBSMT system	No separate training	High, due to the time required for merging
Decoding Time	Low, just as much as a baseline PBSMT system	Very high, due to multiple decoding	High due to increase in model size
Model Size	training corpus size <=2*max(src-pvt,pvt-tgt) corpus same order as PBMST model of this size	No new model created	Blow-up due to the join during merge
Translation Accuracy	could be comparable to cascaded model	taking top-n candidates better than top-1	best method

Translation Accuracies (Case Studies)

Marino & Gispert, 2006

- Catalan-English with Spanish as pivot
- **Cascaded & Synthetic approaches are comparable**

	BLEU	WER	PER
Cat → Eng (cascaded)	0.5147	36.31	27.08
Cat → Eng (synthetic)	**0.5217**	**35.79**	**26.79**
Spa → Eng	0.5470	34.41	25.45
Eng → Cat (cascaded)	**0.4680**	40.66	32.24
Eng → Cat (synthetic)	0.4672	**40.50**	**32.11**
Eng → Spa	0.4714	40.22	31.41

Utiyama & Isahara, 2007

- Various European languages with English as pivot
- **Triangulation is the better than cascading**
- **using top-n(=15) candidates better than top-1 for cascading method**
- The triangulation method is comparable to the direct translation system (>90% of direct system's performance as measured by BLEU)

Source-Target	Direct		Triangulation		Cascading (n=15)		Cascading(n=1)
Spanish–French	35.78	>	32.90 (0.92)	>	29.49 (0.82)	>	29.16 (0.81)
French–Spanish	34.16	>	31.49 (0.92)	>	28.41 (0.83)	>	27.99 (0.82)
German–French	23.37	>	22.47 (0.96)	>	22.03 (0.94)	>	21.64 (0.93)
French–German	15.27	>	14.51 (0.95)	>	14.03 (0.92)	<	14.21 (0.93)
German–Spanish	22.34	>	21.76 (0.97)	>	21.36 (0.96)	>	20.97 (0.94)
Spanish–German	15.50	>	15.11 (0.97)	>	14.46 (0.93)	<	14.61 (0.94)

Augmentation Methods

- Linear Interpolation
- Fillup Interpolation
- Multiple Decoding Paths

Linear Interpolation (Wu & Wang,2009)

- Given *n* models (direct+pivots), combine them to create a single translation model via linear interpolation of models
- Interpolation of phrase translation & lexical probability for PBSMT

$$\phi(\overline{f} \mid \overline{e}) = \sum_{i=0}^{n} \alpha_i \phi_i(\overline{f} \mid \overline{e})$$

$$p_w(\overline{f} \mid \overline{e}, a) = \sum_{i=0}^{n} \beta_i p_{w,i}(\overline{f} \mid \overline{e}, a)$$

where, α_i and β_i are interpolation weights for model *i* for each feature

- Choosing interpolation weights
  - Higher weight to direct model
  - Weighted by BLEU score of standalone systems
  - Tune on development set

Fillup Interpolation (Dabre et al, 2015)

- Back-off scheme
- Define a priority of the models being combined
- Create a single phrase table by choosing entries from the input models in order of priority
- Look into the next model only if an entry is not found in the higher ranked input model

- No modification of probabilities
- Defining the priority of pivots
  - based on translation quality of each individual model
    - Direct system would most likely be first!
  - based on similarity between source/target and pivot languages

Multiple Decoding Paths (MDP) (Nakov & Ng, 2009 ; Dabre et al, 2015)

- Runtime integration
- Decoder searches over all phrase tables for translation options
- Each model will result in its own hypothesis
- The decoder will score each of the hypothesis and select the best one

- Cannot define priority or weighting of the different phrase tables
  - These tend to be ad-hoc anyway
- Makes up for this limitation by allowing multiple models to compete with each other

Comparison of Augmentation Methods

Criteria	Linear Interpolation	Fillup	MDP
Ease of implementation	Easy, tuning the interpolation weights is tricky	Easy	Difficult
Training Time	Tuning time could be enormous	Merging the tables can be done efficiently	No overhead
Decoding Time	No overhead	No overhead	High due to searching over multiple paths
Weighting of Models	Yes	Yes	No
Translation Accuracy	marginal improvement over direct model, may not be statistically significant	performance comparable to linear interpolation	best method, gives significant improvement over direct system

Translation Accuracies (Case Studies) (Dabre et al, 2015)

- Japanese-Hindi translation using various pivots
- Not clear if any of the linear interpolation is better than other
- Performance of Fillup and linear interpolation cannot be distinguished
- **MDP is clearly better than all interpolation schemes**

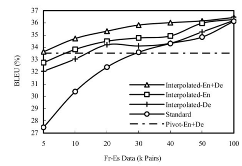
(1): Priority (9:1 ratio for Direct:Bridge table), (2) Priority by BLEU score

Pivot Language	Linear Interpolate (1) With Direct	Linear Interpolate (2) With Direct	Fill Interpolate With Direct	MDP With Direct
1. Direct	33.86			
2. Chinese	34.03	**34.61**	34.31	**35.66**
3. Korean	**34.65**	34.18	34.64	**35.60**
4. Esperanto	**34.63**	**34.55**	35.32	**35.74**

Effect of Multiple Pivots

Fr-Es translation using 2 pivots

Source: Wu & Wang (2007)



Hi ⟷ Ja translation using 7 pivots

Source: Dabre et al (2015)

System	Ja→Hi	Hi→Ja
Direct	33.86	37.47
Direct+best pivot	35.74 (es)	39.49 (ko)
Direct+Best-3 pivots	38.22	41.09
Direct+All 7 pivots	38.42	40.09

- Adding a pivot increases vocabulary coverage
- **Does adding more pivots help?**
- **The answer fortunately is YES!**
- Especially useful when the training corpora are small

What is a good pivot?

- Core concepts
- **What is a good pivot?**
- Addressing language divergences in pivot based SMT

What is a good pivot? (Paul et al, 2013)

- **Supplementary Que**: Is English always a good pivot? Important since English is the *lingua franca* of the world
- A difficult question to answer
- Some *rule-of-thumb guidelines* based on extensive empirical work by Paul etal (2013) on 22 Indo-European & Asian languages

(Indo-European Languages)

Language		Voc	Len	OOV	Order	Unit	Inflection
Danish	DA	26.5k	7.2	1.0	SVO	word	high
German	DE	25.7k	7.1	1.1	mixed	word	high
English	EN	15.4k	7.5	0.4	SVO	word	moderate
Spanish	ES	20.8k	7.4	0.8	SVO	word	high
French	FR	19.3k	7.6	0.7	SVO	word	high
Hindi	HI	33.6k	7.8	3.8	SOV	word	high
Italian	IT	23.8k	6.7	0.9	SVO	word	high
Dutch	NL	22.3k	7.2	1.0	mixed	word	high
Polish	PL	36.4k	6.5	1.1	SVO	word	high
Portuguese	PT	20.8k	7.0	1.0	SVO	word	high
Brazilian Portuguese	PTB	20.5k	7.0	1.0	SVO	word	high
Russian	RU	36.2k	6.4	2.3	SVO	word	high

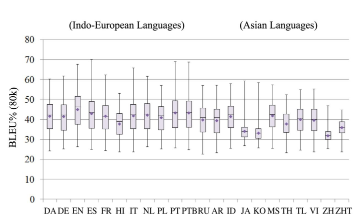
(Asian Languages)

Language		Voc	Len	OOV	Order	Unit	Inflection
Arabic	AR	47.8k	6.4	2.1	VSO	word	high
Indonesian	ID	18.6k	6.8	0.8	SVO	word	high
Japanese	JA	17.2k	8.5	0.5	SOV	none	moderate
Korean	KO	17.2k	8.1	0.8	SOV	phrase	moderate
Malay	MS	19.3k	6.8	0.8	SVO	word	high
Thai	TH	7.4k	7.8	0.4	SVO	none	light
Tagalog	TL	28.7k	7.4	0.7	VSO	word	high
Vietnamese	VI	9.9k	9.0	0.2	SVO	phrase	light
Chinese	ZH	13.3k	6.8	0.5	SVO	none	light
Taiwanese	ZHT	39.5k	5.9	0.6	SVO	none	light

Good diversity in terms of the linguistic phenomena

Is there a single best pivot?

- **There is no single "best" pivot language**
- English is a good pivot in 45.2% (190 out of 230) of the language pairs
- However, 54.8% language pairs chose other pivots



Plots BLEU scores of systems for each pivot

Which pivots are generally good?

(All Languages)

PVT	usage	(%)
EN	232	(50.2)
PT	40	(8.7)
PTB	38	(8.2)
ID	37	(8.0)
MS	36	(7.8)
JA	29	(6.3)
KO	21	(4.5)
ES	19	(4.1)
NL	5	(1.1)
ZH	4	(0.9)
ZHT	1	(0.2)

Among non-English pivots

(Indo-European)

PVT	usage	(%)
PT	40	(36.3)
PTB	32	(29.1)
ES	26	(23.7)
NL	10	(9.1)
DE	1	(0.9)
DA	1	(0.9)

(Asian)

PVT	usage	(%)
ID	28	(31.1)
MS	27	(30.0)
JA	15	(16.6)
KO	12	(13.3)
ZH	4	(4.4)
ZHT	2	(2.2)
VI	1	(1.1)
AR	1	(1.1)

- **Closely related languages are generally good pivots** (Indonesian-Malay, Japanese-Korean, Portuguese-Brazilian Portuguese)
- Portuguese, Brazilian Portuguese best non-English pivots for European languages
- Indonesian, Malay best non-English pivots for European languages

Training Data Size Dependency

- By and large, pivot language for a given language pair is independent of the data size (~86%)
- For the remaining cases, the following trend was observed:
  - For small training data, pivot language related to the source is preferred
  - For larger training data, pivot language related to the target is preferred

$BTEC_{10K}$ PVT	$BTEC_{80K}$ PVT	Language Pair
EN	**ID** (11)	DA-**MS**, ES-**MS**, FR-**MS**, IT-**MS**, PL-**MS**, RU-**MS**, TL-**MS**
JA		KO-**MS**, ZH-**MS**
KO		JA-**MS**
PTB		PT-**MS**
KO	**EN** (9)	JA-**DA**, JA-**DE**, JA-**FR**, JA-**IT**, JA-**NL**, JA-**PL**, JA-**RU**, ZH-**ES**, ZH-**IT**
EN	**KO** (8)	DA-**JA**, ES-**JA**, HI-**JA**
FR		PL-**JA**
ID		VI-**JA**
PT		PTB-**JA**, TL-**JA**
PTB		PT-**JA**
EN	**MS** (3)	DA-**ID**
JA		ZH-**ID**
PTB		PT-**ID**
en	**JA** (3)	FR-**KO**, VI-**KO**
ms		ID-**KO**
JA	**PTB** (2)	ZH-**PT**
MS		ID-**PT**
KO	**PT**	JA-**PTB**

$BTEC_{10K}$ PVT	$BTEC_{80K}$ PVT	Language Pair
ES	EN (18)	RU-IT
FR		IT-JA
ID		TH-ZHT
JA		ZH-TH, ZH-VI
NL		DE-JA
PT		DA-PTB, NL-PTB, FR-JA, NL-JA
PTB		ES-IT, FR-IT, AR-JA, ZHT-IT
ZH		ZHT-TH, ZHT-VI
ZHT		ZH-FR, ZH-TL
EN	ES (2)	FR-ZH
FR		IT-ZH
EN	ID (2)	MS-JA
JA		TH-ZH
KO	JA (2)	ZH-HI, ZH-ZHT
EN	NL (2)	ZH-DE
KO		ZH-AR

Addressing Language Divergence in Pivot-based MT

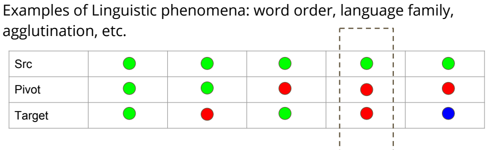
Primary divergence factors affecting translation (Birch, 2008)

- Lexical divergence
- Word order divergence between source and target
- Morphological divergence

- Core concepts
- What is a good pivot?
- **Addressing language divergences in pivot based SMT**

Divergence Scenarios in Pivot-SMT

- Same colour indicates that the languages are not divergent for the linguistic phenomena under consideration
- Examples of Linguistic phenomena: word order, language family, agglutination, etc.



Addressing Word-Order divergence (Patil, Chavan et al, 2015)

Scenario

- Word Order Divergence between source and target language
- **Given a source-pivot and pivot-target lexicalized reordering model, obtain a source-target lexicalized reordering model**
  - For the phrase pairs that are newly added through Phrase Table Triangulation, no reordering information is available
  - Why lexicalized reordering model?: language agnostic and no additional resource requirements
- Use of pivot language to assist the direct translation system

Triangulating Lexicalized Reordering Model

- Lexicalised reordering model contains a reordering table with 6 probability values
- Task is to learn these values in the triangulated table

By looking |||पाहिल्यावर|||$\underbrace{0.2\ 0.2\ 0.6}_{\text{w.r.t Previous Phrase}}\ \underbrace{0.2\ 0.2\ 0.6}_{\text{w.r.t next Phrase}}$

(pAhilyAvara)

Use only the original reordering tables (source→ pivot and pivot→ source) plus a weighting factor which decides how important each entry from the original tables are.

Two way of determining the weighting factor:

- **Heuristic (table-based):** Some heuristics to determine the weighting factors equally among possible reorderings
- **Corpus-driven (count-based):** Determined from the alignments in both the parallel corpora

Case Study

Language Combination	Without Reordering triangulation	With Reordering triangulation
En-Hi-Gu	17.57	**17.67**
En-Hi-Mr	13.17	**13.18**

Table based method

Language Combination	Without Reordering triangulation	With Reordering triangulation
En-Hi-Gu	17.37	**17.71**
En-Hi-Mr	13.11	**13.19**

Count based method

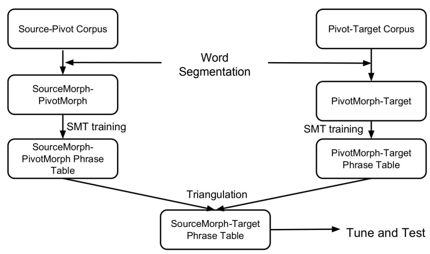
Note: The above are augmented systems (using interpolation) & lexicalized reordering is used

- Table-based method does not always significantly outperform direct reordering system
- Reason: The values of the multiplicative factors have been set heuristically, without consideration to evidence from the data
- Count-based method utilizes evidence from the data to compute the multiplicative factor
- Consistently outperforms direct reordering system

Addressing morphological divergence (More et al, 2015)

Scenario:

- **Agglutinative source language & non-agglutinative target**
- Pivot may/may not be agglutinative
- Use of pivot language to assist the direct translation system



Case Study: Malayalam-Hindi translation

Source: Malayalam (agglutinative)
Target: Hindi (not agglutinative)
Pivots: Bangla, Gujarati, Punjabi (not agglutinative)
Konkani, Marathi, Tamil, Telugu (agglutinative)

System	% BLEU
Direct	16.11
Direct+All Pivot	18.67
Direct (source segmented)	23.35
Direct+All Pivot (source, pivot segmented)	25.51

Effect of Triangulation: Augmentation by pivot improves BLEU Score by 15% over direct system

Effect of Triangulation+Word segmentation: Rise in BLEU score by 58% over direct system

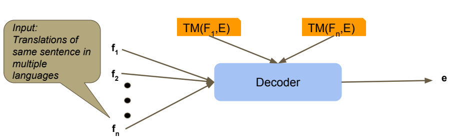
Segmenting both pivot and source is beneficial: Word segmentation on pivot level as well gives BLEU score increase of 4% to 18% over word segmentation at source only, depending on the pivot used

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- **Synergy among multiple languages**
  - Pivot based SMT
  - **Multi-source translation**
- Summary & Conclusion
- Tools & Resources

Multi-source translation

Introduction



- Useful in a scenario where translations are generated in multiple languages
  - EU proceeding, United Nations
- Translations already generated could help subsequent languages:
  - Better word sense disambiguation & other ambiguities
  - Better word order
- Specific case of this scenario: Multiple inputs in the same language which are paraphrases of each other

Model (Och & Ney, 2001)

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}|\mathbf{f}_1^N)\} \\ &= \arg\max_{\mathbf{e}} \{Pr(\mathbf{e}) \cdot Pr(\mathbf{f}_1^N|\mathbf{e})\}\end{aligned} \quad (1)$$

Input sentences are assumed to be independent given the target sentence to simplify modelling

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot \prod_{n=1}^{N} p(\mathbf{f}_n|\mathbf{e})\} \quad (2)$$

Decoding with this scheme is not tractable

- requires enumeration of all target strings
- evaluate permutations from various parts of source string for combination

Solution: Approximations to the decoding objective which make it computationally tractable

Approximate decoding schemes (Och & Ney, 2001)

PROD Model

- **Restrict hypothesis space to the best target sentences from each input sentence**
- This can be done using a standard single source decoder

$$\mathbf{e}_n = \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot p(\mathbf{f}_n|\mathbf{e})\}, \quad n = 1, \ldots, N$$

- For each candidate $\mathbf{e}_n$, the translation model scores all translation models are computed
- The candidates are then scored using the simplified model (2) on previous slide

MAX Model

- Simplifies the decoding objective even further
- **Just chooses the best translation out of the target translation from each decoder**

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \{p(\mathbf{e}) \cdot \max_{n} p(\mathbf{f}_n|\mathbf{e})\} \\ &= \arg\max_{\mathbf{e},n} \{p(\mathbf{e}) \cdot p(\mathbf{f}_n|\mathbf{e})\} \quad .\end{aligned}$$

Limitations

- **Hypothesis space is restricted to a great extent**
- Limited to selecting the best translation from amongst each individual system
- Cannot combine translation options from different language pair models

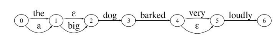
Combining translation options from multiple languages

Output Combination *(Matusov et al, 2006; Schroeder et al, 2009)*

- **Post-processing approach**
- Get top-k translations from each language-pair's model
- **Stitch together a new translation by combining translation fragments from different outputs**
- Rescore the newly composed translation using language model & other features
- Common representation (like ***confusion network***) to represent all outputs for combination

the	ϵ	dog	barked	very	loudly
a	big	dog	barked	ϵ	loudly
sub	*insert*	–	*shift*	*delete*	–

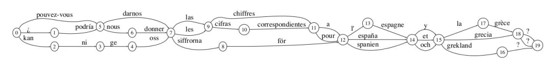
Translation options



Confusion network

Input Combination *(Schroeder et al, 2009)*

- **Select input fragments from different input sentences**
- Create a common *lattice* to represent the multiple inputs
- Input the confusion network to the decoder



Case Study *(Schroeder et al, 2009)*

- **Multi-source translation performs better than single source for even the simplest method, MAX**
- Adding more input languages:
  - no improvement for MAX
  - Improves quality for PROD, input and output combination
- MAX better than PROD for 2 input languages *(Och, Ney 2001)*
- **Output combination is the best method**
- Input combination shows promise

Approach	test2006	test2007
French Only	29.72	30.21
French + Swedish		
MAX	29.86	30.13
LATTICE	29.33	29.97
MULTILATTICE	29.55	29.88
SYSCOMB	31.32	31.77
French + Swedish + Spanish		
MAX	30.18	30.33
LATTICE	29.98	30.45
MULTILATTICE	30.50	30.50
SYSCOMB	33.77	33.87
6 Languages		
MAX	28.37	28.33
LATTICE	30.22	30.91
MULTILATTICE	30.59	30.59
SYSCOMB	35.47	36.03

BLEU scores for English as target language
MAX: Max approach
SysComb: output combination
Lattice & MultiLattice: input combination methods
MultiLattice uses multiple confusion networks

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
- **Summary & Conclusion**
- Tools & Resources

Summary & Conclusion

Let's look back at the questions we started with

- What does it mean to say languages are related?
- Can translation between related languages be made more accurate?
- Can multiple languages help each other in translation?
- Can we reduce resource requirements?

- Universal translation seems difficult. Can we find the right level of linguistic generalization?
- Can we scale to a group of related languages?

- What concepts and tools are required for solving the above questions?

What does it mean to say languages are related?

- Genetic relation → Language Families
- Contact relation → *Sprachbund* (Linguistic Area)
- Linguistic typology → Linguistic Universal
- Orthography → Sharing a script

India as a 'linguistic area'

Exercise

- Are there other notions of relatedness?
- How does relatedness help?

Can we reduce resource requirements?

- Small set of common rules for tasks involving Brahmi-derived scripts:
  - Rule-based transliteration
  - Approximate syllabification
  - Bootstrapping unsupervised transliteration

  Made possible by consistent script principes & systematic design of Unicode encoding
- Common set of source reordering rules for English-Indian languages due to the common canonical word order among Indian languages
- Reduction in parallel corpus requirement due to orthographic similarity :
  - Easily detect cognates, named entities to augment the parallel corpus
  - Translate words not represented in parallel corpus

Can language relatedness of improved translation/transliteration?

- Orthographic Similarity: Properties of Brahmi-derived scripts to improve transliteration
  - Approximate syllabification via vowel segmentation made possible by script properties
  - **There is a lot of potential to harness the scientific design of Indic scripts**
- Lexical & Phonetic Similarity help us do the following:
  - Improve word alignment
  - Translate OOVs
  - Character-oriented SMT
    - **Character-oriented SMT between arbitrary language pairs has shown some promising, may be worth investigating**
- Morphological Similarity: Data sparsity reduction manifests as significant gains in translation accuracy
- Syntactic Similarity: We get a free ride because of similar word order

Can multiple languages help each other?

- Improvement in translation & transliteration performance due to synergy among multiple languages
- Pivot-based translation helps translation by bringing in additional translation options and increasing vocabulary coverage
- Multi-source translation helps translate better by using other languages to reduce linguistic ambiguities during translation
- Related languages contribute most to improvement
- **Bridging divergence gap among languages involved is important**
- What is a good pivot?
  - Related language
  - Morphologically simple
  - English is always an option due to the rich availability of resources involving English
- **Understanding the mechanisms in which various languages interact in a pivot-based setup is an open question**

Key Tools & Concepts

- Language Typology
- Phonetic properties
- Phonetic & Orthographic similarity
- Cognate Identification
- Confusion networks & Word lattices
- Triangulation of translation models
- System combination of SMT output

Related Work that might be of interest

- Study of linguistic typology
- Historical/Comparative linguistics
- Mining bilingual dictionaries and named entities
- Mining parallel corpora
- Word alignment using bridge languages
- Unsupervised bilingual morphological segmentation
- Character-oriented SMT for arbitrary languages
- **Rule-based and Example-based MT in the light of linguistic similarities**

What is the right level of generalization to build an MT system?

Design Goals

- Broad coverage of multiple languages
- Reasonably accurate translation (*indicative translations*)
- Reduce the linguistic resources required

- Universal translation schemes cannot achieve all these goals
- Building customized solutions for every language pair is not feasible

Is a language family or linguistic area a good level of generalization?

Language Relatedness & Translation Accuracy

	Indo-Aryan							Dravidian			
	hin	urd	pan	ben	guj	mar	kok	tam	tel	mal	eng
(A) Phrase based system (S1)											
hin	-	50.30	70.06	36.31	53.29	33.78	36.06	11.36	21.59	10.95	28.15
urd	58.09	-	51.90	26.14	38.92	21.21	25.09	8.13	14.65	7.49	21.00
pan	71.26	44.46	-	30.27	46.24	25.54	29.44	8.96	17.92	7.49	24.01
ben	36.16	24.91	31.84	-	31.24	19.79	23.16	8.88	13.18	8.62	18.34
guj	53.09	34.77	47.60	29.35	-	26.99	29.63	9.95	16.57	7.97	19.58
mar	41.66	25.08	34.75	23.68	33.84	-	27.44	8.34	12.02	7.25	15.87
kok	38.54	25.54	33.53	24.61	31.44	23.69	-	7.96	13.40	8.05	16.92
tam	21.79	15.65	19.32	14.77	17.28	11.10	14.17	-	9.30	6.41	10.90
tel	27.20	19.03	25.14	16.87	22.22	13.47	16.98	7.29	-	6.58	12.09
mal	14.50	10.27	12.53	10.01	10.99	7.01	9.36	4.67	6.25	-	8.36
eng	26.53	18.07	22.86	14.85	17.36	10.17	13.01	4.17	6.43	4.85	-

Translation Accuracy vis-a-vis Language Families

- Clear partitioning of translation pairs by language family pairs, based on translation accuracy
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult
- Language families: The right level of generalization for building SMT systems

Is the clear partitioning indicative that the language family forms a good unit of abstraction?

Where are we?

- Motivation
- Language Relatedness
- A Primer to SMT
- Leveraging Orthographic Similarity for transliteration
- Leveraging linguistic similarities for translation
  - Leveraging Lexical Similarity
  - Leveraging Morphological Similarity
  - Leveraging Syntactic Similarity
- Synergy among multiple languages
- Summary & Conclusion
- Tools & Resources

Tools & Resources

Language & Variation

- Ethnologue: Catalogue of all the world's living languages (www.ethnologue.com)
- World Atlas of Linguistic Structures: Large database of structural (phonological, grammatical, lexical) properties of languages (wals.info)
- Comrie, Polinsky & Mathews. *The Atlas of Languages: The Origin and Development of Languages Throughout the World*
- Daniels & Bright. *The World's Writing systems*.

Tools

- Pivot-based SMT: https://github.com/tamhd/MultiMT
- System Combination: MEMT
- Moses contrib has tools for combining phrase tables
- Moses can take confusion network as input
- Multiple Decoding Paths is implemented in Moses

Machine Translation & Transliteration Resources @ IIT Bombay

Software

CFILT Pre-Order

- URL: http://www.cfilt.iitb.ac.in/~moses/download/cfilt_preorder/register.html
- Rule-based Source reordering system for English to Indian Language translation
- Python and command line interfaces
- In progress: parallelization of the Python API
- Shows improvement across many English-IL systems
- GPL licensed

	Indo-Aryan							Dravidian			
	hin	urd	pan	ben	guj	mar	kok	tam	tel	mal	eng
(A) Phrase based system (S1)											
eng	26.53	18.07	22.86	14.85	17.36	10.17	13.01	4.17	6.43	4.85	-
(B) Phrase based system with source reordering: generic rules (S2)											
eng	29.63	20.42	26.06	16.85	20.11	11.46	15.01	4.97	7.83	5.53	-
(C) Phrase based system with source reordering: Hindi-adapted rules (S3)											
eng	30.86	21.54	27.52	18.20	21.33	12.68	15.73	5.09	8.29	5.68	-

- Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference. 2014.
- R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya and M. Sasikumar, *Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation*, IJCNLP. 2008.

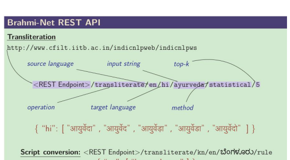
METEOR-Indic

- METEOR for 17 Indian languages
- Supports the following matching modules:
  - Synonyms (using IndoWordnet)
  - Stem (using a Trie based matcher)
- Available on request
  - You need access to IndoWordnet data
  - Hindi/Marathi/Sanskrit wordnets are freely available for research use

- Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference. 2014.
- Anoop Kunchukuttan, Ratish Puduppully, Rajen Chatterjee, Abhijit Mishra, Pushpak Bhattacharyya. 2014. *The IIT Bombay SMT System for ICON 2014 Tools Contest* . NLP Tools Contest at ICON 2014. 2014.

Transliteration Tools (BrahmiNet)

- Script Conversion among Indic scripts (16 languages)
- Romanization for Indic scripts (16 languages)
- Machine Transliteration among 18 languages
- Available as REST Web Service
- Documentation: http://www.cfilt.iitb.ac.in/brahminet/static/rest.html
- Planned: Python client in Indic NLP Library
- Script conversion & romanization can also be accessed offline using the Indic NLP library



Anoop Kunchukuttan, Ratish Puduppully , Pushpak Bhattacharyya, *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent* , Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations **(NAACL 2105)** . 2015.

Indic NLP Library

- Library of NLP components for Indian languages
- Easy to install and use
- Generic framework for Indian languages
- Website: http://anoopkunchukuttan.github.io/indic_nlp_library/
- Documentation: http://indic-nlp-library.readthedocs.org

Language Support

Indo-Aryan			Dravidian	Others
Assamese (asm)	Marathi (mar)	Sindhi (snd)	Kannada (kan)	English (eng)
Bengali (ben)	Nepali (nep)	Sinhala (sin)	Malayalam (mal)	
Gujarati (guj)	Odia (ori)	Sanskrit (san)	Telugu (tel)	
Hindi/Urdu (hin/urd)	Punjabi (pan)	Konkani (kok)	Tamil (tam)	

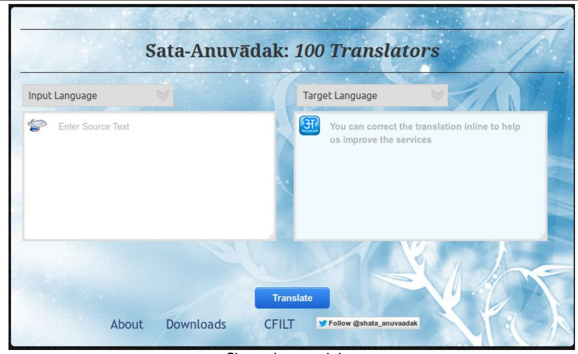
Tasks

Monolingual	Indo-Aryan													Dravidian			
	san	hin	urd	pan	nep	snd	asm	ben	ori	guj	mar	kok	sin	kan	tel	tam	mal
Script Information	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Normalization	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Word segmentation	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓
Romanization (ITRANS)	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ITRANS to Script	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Bilingual

- **Script Conversion:** Amongst the above mentioned languages, except Urdu and English
- **Transliteration:** Amongst the 18 above mentioned languages
- **Translation:** Amongst these 10 languages: (hin, urd, pan, ben, guj, mar, kok, sin, kan, tel, tam, mal) + English

Online Systems

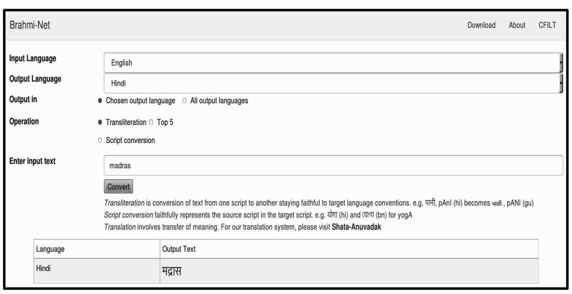


Shata-Anuvaadak

http://www.cfilt.iitb.ac.in/indic-translator/

110 language pairs

English, 7 Indo-Aryan & 3 Dravidian languages



Brahmi-Net

http://www.cfilt.iitb.ac.in/brahminet/

306 language pairs

English, 13 Indo-Aryan & 7 Dravidian languages

Resources

Brahmi-Net Transliteration Corpus

- 1.6 million word pairs among 10 Indian languages (+English)
- Mined from the ILCI corpus
- URL: http://www.cfilt.iitb.ac.in/brahminet/static/register.html
- License: Creative Common Attribution-NonCommercial (CC BY-NC)

Anoop Kunchukuttan, Ratish Puduppully , Pushpak Bhattacharyya, *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent* , Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations **(NAACL 2105)** . 2015.

	Indo-Aryan							Dravidian			
	hin	urd	pan	ben	guj	mar	kok	tam	tel	mal	eng
hin	-	21185	40456	26880	29554	13694	16608	9410	17607	10519	10518
urd	21184	-	23205	11379	14939	9433	9811	4102	5603	3653	5664
pan	40459	23247	-	25242	29434	21495	21077	7628	15484	8324	8754
ben	26853	11436	25156	-	33125	26947	26694	10418	18303	11293	7543
guj	29550	15019	29434	33166	-	39633	35747	12085	22181	11195	6550
mar	13677	9523	21490	27004	39653	-	31557	10164	18378	9758	4878
kok	16613	9865	21065	26748	35768	31556	-	9849	17599	9287	5560
tam	9421	4132	7668	10471	12107	10148	9838	-	12138	10931	3500
tel	17649	5680	15598	18375	22227	18382	17409	12146	-	12314	4433
mal	10584	3727	8406	11375	11249	9788	9333	10926	12369	-	3070
eng	10513	5609	8751	7567	6537	4857	5521	3549	4371	3039	-

Diverse types of transliterations

Category	Example	Extended ITRANS transliteration
Named Entities	(ಅಂಧೇರಿ, अंधेरी) (ಅಕ್ಬರ್, अकबर)	(aMdherI,aMdherI) (akabara,akabara)
Spelling variations	(telephone ,टेलीफोन/टेलिफोन) (Belgaum , बेलगाँव/बेलगाम) (फेब्रुवारी, फरवरी)	(, TelIphona/Teliphona) (, belagA.Nva/belagAma) (phebruvArI, pharavarI)
Tatsam words[5]	(अहंकार,അഹങ്കാരം) (करुणा,കരുണ) (चक्र,ചക്രം)	(aha.nkAra,aha NkAra.n) (karuNA,karuNa) (cakra,cakra.n)
English Loan words	(syphilis, सिफिलिस) (telephone, टेलिफोन) (কাউন্সিলিং,counselling)	(,siphilisa) (, Teliphona) (kAunsili.n,)
Indian origin words	(tandoori, तंदूरी) (avatar,अवतार) (yoga, योगा)	(tandoori, ta.MdUrI) (avatar,avatAra) (yoga, yogA)
ΨSound shifts	(केरळ, केरल)	(keraL, keral)
Cognates	(अंधळेपणा, अंधेपन) (कसे, कैसे) (गाढव, गधा) (பக்தர்கள், भक्तगण)	(aMdhLepaNa, aMdhepan) (kase, kaise) (gaDhav, gadha) (paktarkaL, bhaktagaN)
Script differences	(ஏரோபிக்ஸ் , एरोबिक्स) (ஏரோபிக்ஸ் , गंगोत्री) (अमृतस,അമൃതസർ)	(eropiks,erobiks) (ka~Nkotari,gaMgotrI) (amRitasara,amRitasar)

Xlit-Crowd: Hindi-English Transliteration Corpus

- The corpus contains transliteration pairs for Hindi-English
- Obtained via crowdsourcing using Amazon Mechanical Turk by asking workers to transliterate Hindi words into Roman script
- The source words for the task came from NEWS 2010 shared task corpus
- Size: 14919 transliteration pairs

Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah, Pushpak Bhattacharyya. *When Transliteration Met Crowdsourcing : An Empirical Study of Transliteration via Crowdsourcing using Efficient, Non-redundant and Fair Quality Control* . Language and Resources and Evaluation Conference (LREC 2014). 2014.

Shata-Anuvaadak Resources

- PBSMT translation models for 110 language pairs
- Language Models for 11 language pairs
- These have been built from the ILCI corpus
- ILCI corpus can be requested from TDIL (http://www.tdil-dc.in)
- If unavailable, these trained models can directly be used
- License: Creative Common Attribution-NonCommercial CC BY-NC

URL: http://www.cfilt.iitb.ac.in/~moses/shata_anuvaadak/register.html

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference. 2014.

Acknowledgments

- Prof. Pushpak Bhattacharyya
- Prof. Malhar Kulkarni
- Rohit More
- Harshad Chavan
- Deepak Patil
- Raj Dabre
- Abhijit Mishra
- Rajen Chatterjee
- Ritesh Shah
- Ratish Puduppully
- Arjun Atreya
- Aditya Joshi
- Rudramurthy V
- Girish Ponkiya

… and everyone at the Center for Indian Language Technology

Thank You!

Questions?

References

- Anvita Abbi. *Languages of India and India and as a Linguistic Area*. 2012. Retrieved November 15, 2015, from http://www.andamanese.net/Languages of India and India as a linguistic area.pdf
- Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. *Statistical machine translation*. Technical report, Johns Hopkins University. 1999
- Shane Bergsma, Grzegorz Kondrak. *Alignment-based discriminative string similarity*. Annual meeting-Association for Computational Linguistics. 2007.
- N. Bertoldi, M. Barbaiani, M. Federico, R. Cattoni. *Phrase-based statistical machine translation with pivot languages*. IWSLT. 2008.
- Alexandra Birch, Miles Osborne, and Philipp Koehn. *Predicting success in machine translation*. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
- Peter Daniels and William Bright. *The world's writing systems*. Oxford University Press, 1996.
- Peter Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. *The mathematics of statistical machine translation: Parameter estimation*. Computational linguistics. 1993.
- Michael Covington. *An algorithm to align words for historical comparison*. Computational linguistics. 1996.
- Raj Dabre, Fabrien Cromiers, Sadao Kurohashi, and Pushpak Bhattacharyya. *Leveraging small multilingual corpora for SMT using many pivot languages*. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015.
- Adri`a De Gispert, Jose B Marino. *Catalan-english statistical machine translation without parallel corpus: bridging through spanish*. In Proc. of 5th International Conference on Language Resources and Evaluation (LREC). 2006.
- Nadir Durrani, Hassan Sajjad, Hieu Hoang and Philipp Koehn. *Integrating an unsupervised transliteration model into statistical machine translation*. EACL. 2014.

References

- Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. *Hindi-to-Urdu machine translation through transliteration*. In Proceedings of the 48th Annual meeting of the Association for Computational Linguistics. 2010.
- Nadir Durrani, Barry Haddow, Phillip Koehn, Kenneth Heafield. *Edinburgh's phrase-based machine translation systems for WMT-14*. Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation. 2014.
- Halvor Eifring, Bøyesen Rolf Theil. *Linguistics for students of Asian and African languages.* Institutt for østeuropeiske og orientalske studier. 2005. Retrieved November 15 2015, from https://www.uio.no/studier/emner/hf/ikos/EXFAC03-AAS/h05/larestoff/linguistics/
- Murray Emeneau. *India as a linguistic area*. Language. 1956.
- Kenneth Heafield, Alon Lavie. *Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme.* The Prague Bulletin of Mathematical Linguistics. 2010.
- Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. *Automatic identification of cognates and false friends in French and English.* Proceedings of the International Conference Recent Advances in Natural Language Processing. 2005.
- Mitesh Khapra, A. Kumaran and Pushpak Bhattacharyya. *Everybody loves a rich cousin: An empirical study of transliteration through bridge languages*. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2010.
- Alexandre Klementiev, Dan Roth. *Weakly supervised named entity transliteration and discovery from multilingual comparable corpora.* Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 2006.
- Philipp Koehn. *Statistical machine translation*. Cambridge University Press. 2009.
- Greg Kondrak. *Cognates and word alignment in bitexts*. MT Summit. 2005.

References

- Grzegorz Kondrak. *A new algorithm for the alignment of phonetic sequences*. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. 2000.
- Greg Kondrak, Daniel Marcu and Kevin Knight. *Cognates can improve statistical translation models*. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2003.
- S. Kumar, Och, F. J., Macherey, W. *Improving word alignment with bridge languages*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007.
- A. Kumaran, Mitesh M. Khapra, and Pushpak Bhattacharyya. *Compositional Machine Transliteration.* ACM Transactions on Asian Language Information Processing. 2010.
- Anoop Kunchukuttan, Ratish Puduppully, and Pushpak Bhattacharyya. *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent.* Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2015.
- Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Sata-Anuvadak: Tackling Multiway Translation of Indian Languages*. Language Resources and Evaluation Conference. 2014.
- G. Mann, David Yarowsky. *Multipath translation lexicon induction via bridge languages.* In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. 2001.
- Evgeny Matusov, Nicola Ueffing, and Hermann Ney. *Computing Consensus Translation for Multiple Machine Translation Systems Using Enhanced Hypothesis Alignment.* EACL. 2006.
- Dan Melamed. *Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons*. Third Workshop on Very Large Corpora. 1995.

References

- Dan Melamed. *Models of translational equivalence among words*. Computational Linguistics. 2000.
- Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura. *Improving Pivot Translation by Remembering the Pivot*. Association for Computational Linguistics. 2015.
- Robert Moore. *A discriminative framework for bilingual word alignment.* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.
- Rohit More. *Pivot based Statistical Machine Translation*. Master's Thesis. IIT Bombay. 2015.
- Rohit More, Anoop Kunchukuttan, Raj Dabre, Pushpak Bhattacharyya. *Augmenting Pivot based SMT with word segmentation*. International Conference on Natural Language Processing. 2015.
- Preslav Nakov, Hwee Tou Ng. *Improving statistical machine translation for a resource-poor language using related resource-rich languages*. Journal of Artificial Intelligence Research. 2012.
- Preslav Nakov, and Jörg Tiedemann. *Combining word-level and character-level models for machine translation between closely-related languages*. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012.
- Preslav Nakov, Hwee Tou Ng. *Improved statistical machine translation for resource-poor languages using related resource-rich languages*. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009.
- Franz Och and Hermann Ney. *Statistical multi-source translation*. In Proceedings of MT Summit VIII. Machine Translation in the Information Age , MT Summit. 2001.
- Franz Och, and Hermann Ney. *A systematic comparison of various statistical alignment models*." Computational linguistics. 2003.
- Raj Nath Patel, Rohit Gupta, and Prakash B. Pimpale. *Reordering rules for English-Hindi SMT*. HYTRA. 2013.
- Deepak Patil, Harshad Chavan and Pushpak Bhattacharyya. *Triangulation of Reordering Tables: An Advancement Over Phrase Table Triangulation in Pivot-Based SMT*. International Conference on Natural Language Processing. 2015.

References

- Michael Paul, Andrew Finch, and Eiichrio Sumita. *How to choose the best pivot language for automatic translation of low-resource languages*. ACM Transactions on Asian Language Information Processing (TALIP). 2013.
- R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya and M. Sasikumar, *Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation*, International Joint Conference on NLP. 2008.
- E. Ristad, P. Yianilos. *Learning string-edit distance.* IEEE Trans. Pattern Anal. Mach. Intell., 20(5):522–532, 1998.
- Hassan Sajjad, Alexander Fraser, and Helmut Schmid. *A statistical model for unsupervised and semi-supervised transliteration mining.* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012.
- J. Schroeder, Cohn, T., and Koehn, P. *Word lattices for multi-source translation*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. 2009.
- Anil Kumar Singh. *A Computational Phonetic Model for Indian Language Scripts.* In proceedings of Constraints on Spelling Changes: Fifth International Workshop on Writing Systems. 2006.
- Harshit Surana and Anil Kumar Singh. *A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages.* In proceedings of the Third International Joint Conference on Natural Language Processing. 2008.
- R. Sinha, Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., and Jain, A.. *ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages.* In IEEE International Conference on Systems, Man and Cybernetics. 1995.
- David Steele, Lucia Specia. *WA-Continuum: Visualising Word Alignments across Multiple Parallel Sentences Simultaneously*. ACL-IJCNLP. 2015.
- Karumuri Subbarao. *South Asian languages : a syntactic typology*. Cambridge University Press. 2012.
- Anil Kumar Singh and Harshit Surana. *Multilingual Akshar Based Transducer for South and South East Asian Languages which Use Indic Scripts*. In Proceedings of the Seventh International Symposium on Natural Language Processing. Pattaya, Thailand. 2007.

References

- Ben Taskar, Simon Lacoste-Julien, and Dan Klein. *A discriminative matching approach to word alignment*. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2005.
- Sarah Thomason. *Linguistic Areas and Language History*. Studies in Slavic and General Linguistics. 2000.
- Jorge Tiedemann. *Character-based PSMT for closely related languages*. In Proceedings of the 13th Annual Conference of the European Association for Machine Translation. 2009.
- Jörg Tiedemann. *Character-based pivot translation for under-resourced languages and domains*. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012.
- Raghavendra Udupa, Mitesh M Khapra. *Transliteration equivalence using canonical correlation analysis*. Advances in Information Retrieval. 2010.
- Masao Utiyama, Hitoshi Isahara. A *comparison of pivot methods for phrase-based statistical machine translation.* In HLT-NAACL, pages 484–491, 2007.
- D. Vilar, Peter, J.-T., & Ney, H.. *Can we translate letters?*. In Proceedings of the Second Workshop on Statistical Machine Translation. 2007.
- Robert Wagner, Michael J. Fischer. *The string-to-string correction problem*. Journal of the ACM. 1974.
- Haifeng Wang, Hua Wu, and Zhanyi Liu. *Word alignment for languages with scarce resources using bilingual corpora of other language pairs*. COLING-ACL. 2006.
- Hua Wu, Haifeng Wang. *Pivot language approach for phrase-based statistical machine translation.* Machine Translation. 2007.
- Robert Östling. *Bayesian word alignment for massively parallel texts*. 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014.