Statistical Machine Translation between Related Languages

Pushpak Bhattacharyya
Indian Institute of Technology Bombay
pb@cse.iitb.ac.in

Anoop Kunchukuttan
Indian Institute of Technology Bombay
anoopk@cse.iitb.ac.in

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

NAACL 2016 Tutorial

San Diego, California

12th June 2016

You can download the slides from: https://www.cse.iitb.ac.in/~anoopk/publications/presentations/naacl-2016-tutorial.pdf

Tutorial Outline

- Introduction & Motivation
- Language Relatedness
- Translation within related languages
- Translation from related languages to another language
- Summary

Tutorial Outline

- **Introduction & Motivation**
- Language Relatedness
- Translation within related languages
- Translation from related languages to another language
- Summary

Parallel Corpus

A boy is sitting in the kitchen	Un garçon est assis dans la cuisine
A boy is playing tennis	Un garçon joue au tennis
A boy sitting on a round table	Un garçon assis sur une table ronde
Some men are watching tennis	Certains hommes regardent le tennis
A girl is holding a black book	Une jeune fille tient un livre noir
Two men are watching a movie	Deux hommes regardent un film
A woman is reading a book	Une femme est en train de lire un livre
A woman is sitting in a red car	Une femme est assise dans une voiture rouge

+

Machine Learning

Lets begin with a simplistic view of Statistical Machine Translation (SMT) !!!

Parallel Corpus

A boy is **sitting** in the kitchen	**Un garçon** est **assis** dans la cuisine
A boy is playing **tennis**	**Un garçon** joue au **tennis**
A boy **sitting** on a round table	**Un garçon** **assis** sur une table ronde
Some men **are watching** **tennis**	Certains hommes **regardent** le **tennis**
A girl is holding a black book	Une jeune fille tient un livre noir
Two men **are watching** a movie	Deux hommes **regardent** un film
A woman is reading a book	Une femme est en train de lire un livre
A woman is **sitting** in a red car	Une femme est **assise** dans une voiture rouge

+

Machine Learning

- Learn word/phrase alignments

Lets begin with a simplistic view of Statistical Machine Translation (SMT) !!!

Parallel Corpus

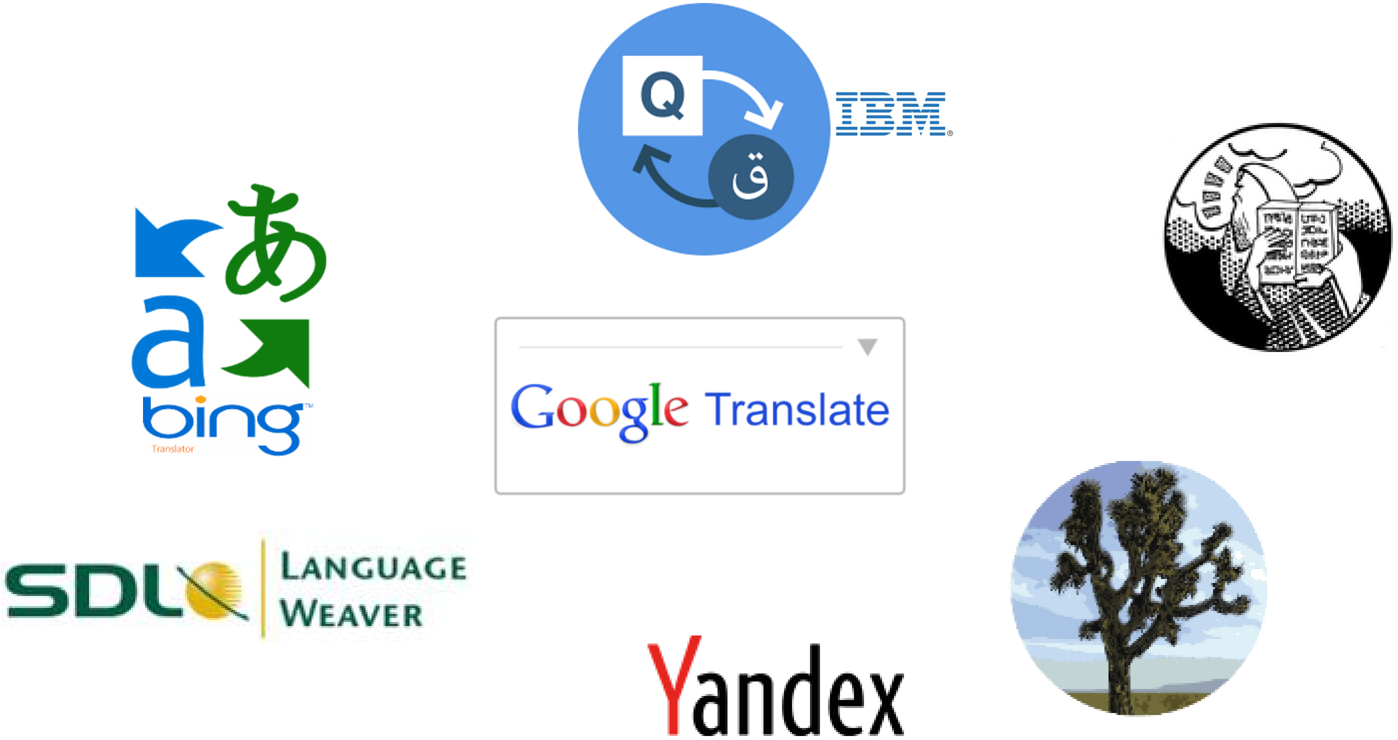
A boy is **sitting** in the kitchen	**Un garçon** est **assis** dans la cuisine
A boy is playing **tennis**	**Un garçon** joue au **tennis**
A boy sitting on a round table	Un garçon **assis** sur une table ronde
Some men **are watching tennis**	Certains hommes **regardent** le **tennis**
A girl is holding a black **book**	Une jeune fille tient un **livre** noir
Two men **are watching** a movie	Deux hommes **regardent** un film
A woman is reading a book	Une femme est en train de lire un livre
A woman is **sitting** in a red car	Une femme est **assise** dans une voiture rouge

+

Machine Learning

- Learn word/phrase alignments
- Learning to reorder

Lets begin with a simplistic view of Statistical Machine Translation (SMT) !!!



SMT is by far the most popular machine translation paradigm

Why is SMT so popular?

… because it is a language independent technology

What do we mean by language independent technology?

“If technology developed for one language can be ported to another merely by amassing appropriate training data in the second language, then the effort put into the development of the technology in the first language can be leveraged to more efficiently create technology for other languages.”

- Emily Bender (2011)

Indeed, by the above definition, SMT is a language independent technology, but….

“If technology developed for one language can be ported to another merely by amassing appropriate training data in the second language, then the effort put into the development of the technology in the first language can be leveraged to more efficiently create technology for other languages.”

- Emily Bender (2011)

but….need to focus on two practical considerations:

"If technology developed for one language can be ported to another merely by amassing appropriate training data in the second language, then the effort put into the development of the technology in the first language can be leveraged to more efficiently create technology for other languages."

- Emily Bender (2011)

but….need to focus on two practical considerations:

- *Not just ported, it should work well!!*
- *How much is 'appropriate' ?*

Even though in theory SMT is language independent, in practice the situation is different ….

HTER	**assessment**	**language pairs and domains**
0%		
	publishable	*French-English restricted domain*
10%		*French-English technical document localization*
	editable	*French-English news stories*
20%		
		English-German news stories
30%	*gistable*	*English-Czech open domain*
40%	*triagable*	
50%		

Source: Philip Koehn, Course slides

Very few languages have high quality SMT systems!!

Lets consider the case of English → Malayalam SMT to understand a few reasons for this ….

Malayalam is a <u>highly agglutinative</u>, predominantly <u>S-O-V</u> language

harder reordering problem

Even if it does not rain

mazhA	p.eyyutil.e~Ngillu.m
rain_NN	rain_VB+not+even_if

too many word forms

leads to data sparsity
(not enough counts for all word forms)

bad word/phrase alignments

<u>Solution</u>

- *Add more parallel data*
- *More linguistic processing (morphological analysis, parsing, etc.)*

Not possible for all languages

A more practical definition of language independent technology should include:

- *~~appropriate~~ less or reusable data*
- *~~appropriate~~ less or portable linguistic resources*

Obviously, this cannot be achieved when porting SMT to arbitrary language pairs

But can this be achieved for some language pairs?

- *Yes, for "related" languages*

The focus of this tutorial

Lets consider the case of Marathi → Hindi SMT to motivate this ….

What’s so special about this language pair

Related by evolution

Belong to the same language family
(Indo-Aryan branch of the IE language family)

Related by contact

Constant exchange between these languages
(Both are spoken in the Indian subcontinent)

… leading to linguistic similarities and prior knowledge that can be used

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

En: On the occasion of India’s Independence day, a program was organized in American city of Los Angeles

India+of	भारताच्या
Independence_day+on_occasion_of	स्वातंत्र्यदिनानिमित्त
America_in	अमेरिकेतील
Los	लॉस
Angeles	एन्जल्स
city+in	शहरात
program	कार्यक्रम
organized	आयोजित
+verbalizer	करण्यात
come+past	आला

En: On the occasion of India’s Independence day, a program was organized in American city of Los Angeles

India	भारता
+of	च्या
Independence	स्वातंत्र्य
Day	दिना
+on_occasion_of	निमित्त
America	अमेरिके
in	तील
Los	लॉस
Angeles	एन्जल्स
city	शहरा
in	त
program	कार्यक्रम
organized	आयोजित
+verbalizer	करण्यात
come+past	आला

1. Segment the Marathi input

En: On the occasion of India's Independence day, a program was organized in American city of Los Angeles

1. Segment the Marathi input
2. Transliterate Named Entities

India	भारता	भारत
+of	च्या	
Independence	स्वातंत्र्य	
Day	दिना	
+on_occasion_of	निमित्त	
America	अमेरिके	अमरीका
in	तील	
Los	लॉस	लॉस
Angeles	एन्जल्स	एन्जल्स
city	शहरा	
in	त	
program	कार्यक्रम	
organized	आयोजित	
+verbalizer	करण्यात	
come+past	आला	

En: On the occasion of India’s Independence day, a program was organized in American city of Los Angeles

India	भारता	भारत
+of	च्या	
Independence	स्वातंत्र्य	स्वतंत्रता
Day	दिना	
+on_occasion_of	निमित्त	
America	अमेरिके	अमरीका
in	तील	
Los	लॉस	लॉस
Angeles	एन्जल्स	एन्जल्स
city	शहरा	शहर
in	त	
program	कार्यक्रम	कार्यक्रम
organized	आयोजित	आयोजित
+verbalizer	करण्यात	
come+past	आला	

1. Segment the Marathi input
2. Transliterate Named Entities
3. Transliterate Cognates and Loan words

En: On the occasion of India’s Independence day, a program was organized in American city of Los Angeles

India	भारता	भारत
+of	च्या	
Independence	स्वातंत्र्य	स्वतंत्रता
Day	दिना	दिवस
+on_occasion_of	निमित्त	
America	अमेरिके	अमरीका
in	तील	
Los	लॉस	लॉस
Angeles	एन्जल्स	एन्जल्स
city	शहरा	शहर
in	त	
program	कार्यक्रम	कार्यक्रम
organized	आयोजित	आयोजित
+verbalizer	करण्यात	किया
come+past	आला	

1. Segment the Marathi input
2. Transliterate Named Entities
3. Transliterate Cognates and Loan words
4. Some more loan words

En: On the occasion of India's Independence day, a program was organized in American city of Los Angeles

India	भारता	भारत
+of	च्या	के
Independence	स्वातंत्र्य	स्वतंत्रता
Day	दिना	दिवस
+on_occasion_of	निमित्त	के [ ] पर
America	अमेरिके	अमरीका
in	तील	के
Los	लॉस	लॉस
Angeles	एन्जल्स	एन्जल्स
city	शहरा	शहर
in	त	में
program	कार्यक्रम	कार्यक्रम
organized	आयोजित	आयोजित
+verbalizer	करण्यात	किया
come+past	आला	

1. Segment the Marathi input
2. Transliterate Named Entities
3. Transliterate Cognates and Loan words
4. Some more loan words
5. Translate function words

En: On the occasion of India’s Independence day, a program was organized in American city of Los Angeles

India	भारता	भारत
+of	च्या	के
Independence	स्वातंत्र्य	स्वतंत्रता
Day	दिना	दिवस
+on_occasion_of	निमित्त	के अवसर पर
America	अमेरिके	अमरीका
in	तील	के
Los	लॉस	लॉस
Angeles	एन्जल्स	एन्जल्स
city	शहरा	शहर
in	त	में
program	कार्यक्रम	कार्यक्रम
organized	आयोजित	आयोजित
+verbalizer	करण्यात	किया
come+past	आला	गया

1. Segment the Marathi input
2. Transliterate Named Entities
3. Transliterate Cognates and Loan words
4. Some more loan words
5. Translate function words
6. Translate remaining **content words**

Why is SMT between Marathi-Hindi different from English-Malayalam?

Machine Learning

- Learn word/phrase alignments
- Learning to reorder

Almost One-One correspondence between words (cognates, loan words, function words)
Transformations at the sub-word level
Level of representation different

They have the same basic word order
The reordering problem is almost non-existent
No parsing is required

Learning at this level requires lesser data

What language divergences still have to be resolved?

“almost” one-to-one correspondence

- *Function words ←→ suffixes e.g. Hindi ←→ Marathi*
- *Function word mappings may not be unique*

1) ghara + ca (of) → ghar + ka (of)
ghara + tIla (in) → ghar + ka (of)

2) hi: raama ko aama pasanda hai
bn: raamera aama pachanda aache

Still need to resolve ambiguity for some content words

- *Translations aren’t orthographically similar: hair: kesa*
- *False Friends: pAnI, panI*

Most translation requirements also involves related languages

Between related languages

Related languages ⇐⇒ Link languages (English, French, Spanish, Hindi, etc.)

Focus of this tutorial:

- *Define relatedness between languages*
- *Exploit relatedness between languages for SMT*
  - *Between related languages*
  - *Between a bunch of related languages and another language*

Tutorial Outline

- Introduction & Motivation
- **Language Relatedness**
- Translation within related languages
- Translation from related languages to another language
- Summary

Let’s start by understanding …

Language Relatedness

How are languages related?

- Genetic Relation → Language Families
- Contact Relation → Linguistic Area
- Linguistic Typology → Linguistic Universal

How are languages related?

- Genetic Relation → Language Families
- Contact Relation → Linguistic Area
- Linguistic Typology → Linguistic Universal

What does language relatedness imply for MT?

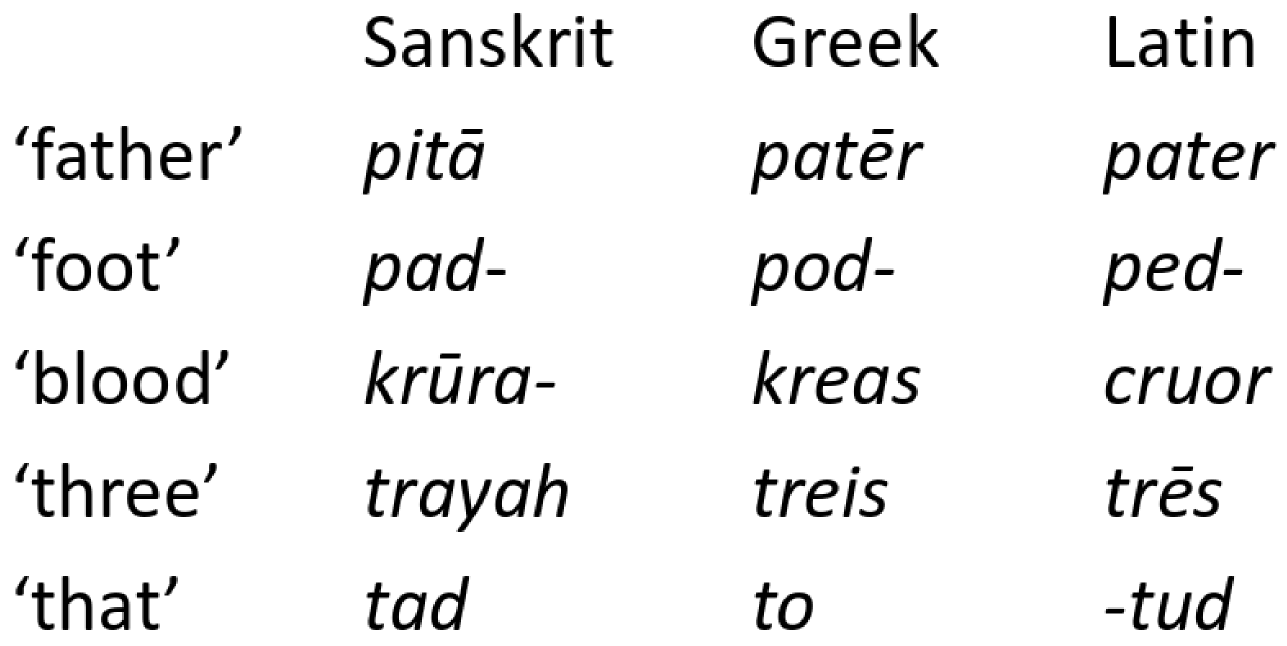
- Cognates (words of the same origin)
- Similar phoneme set, makes transliteration easier
- Similar grammatical properties
  - morphological and word order symmetry makes MT easier
- Cultural similarity leading to shared idioms and multiwords
  - **hin:** दाल में कुछ काला होना (*dAla me.n kuCha kAlA honA)*
  - **guj:** दाळ मा काईक काळु होवु (*dALa mA kAIka kALu hovu)*

Literal meaning: *something black in the lentils*

Idiomatic meaning: *something fishy*

Language Families

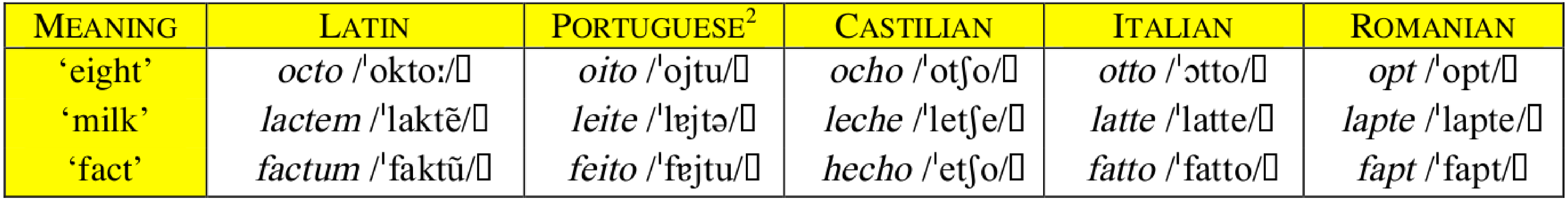
*Group of languages related through descent from a common ancestor, called the **proto-language** of that family*



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

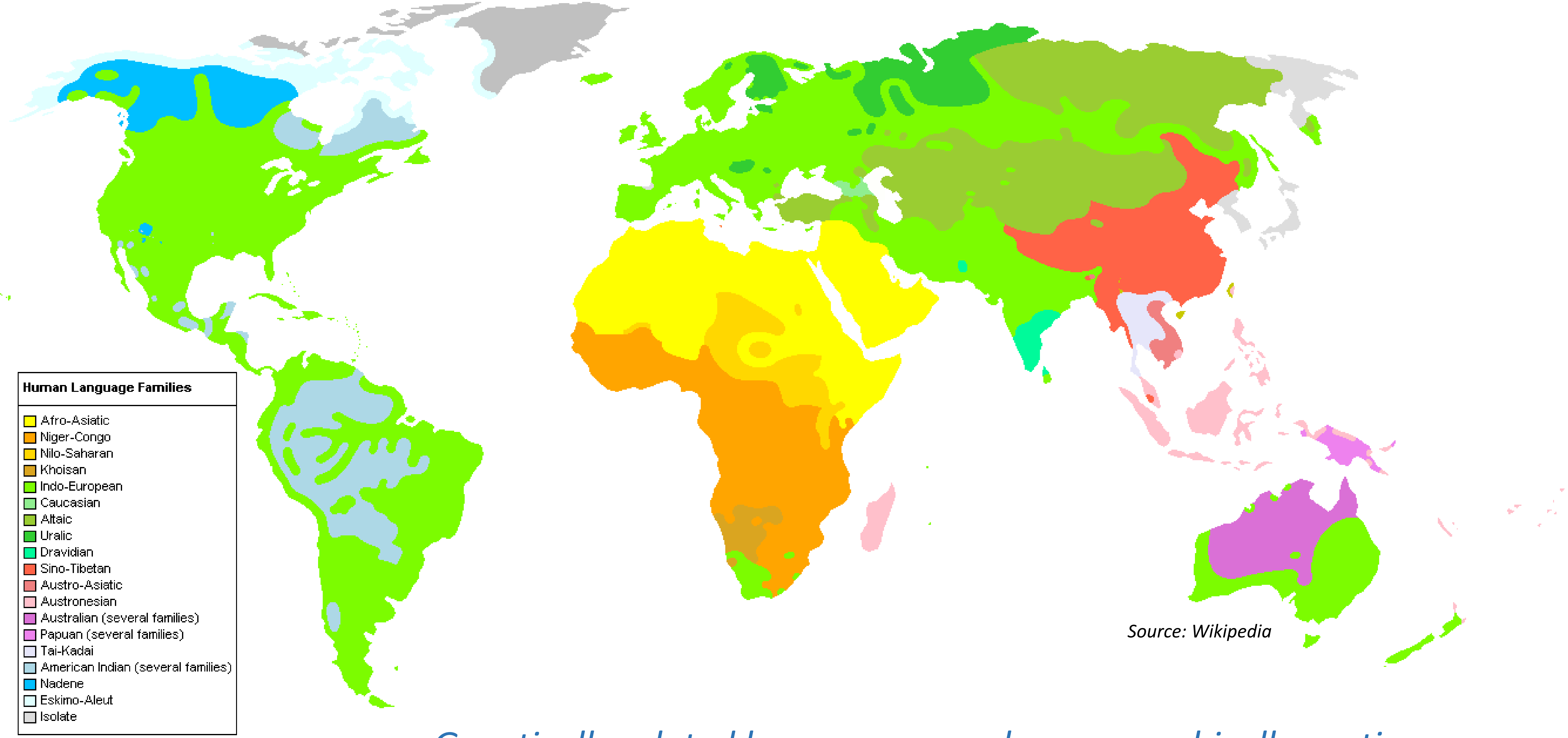
Basis of classification

Regularity of sound change is the basis of studying genetic relationships



MEANING	LATIN	PORTUGUESE[2]	CASTILIAN	ITALIAN	ROMANIAN
‘eight’	*octo* /ˈoktoː/	*oito* /ˈojtu/	*ocho* /ˈotʃo/	*otto* /ˈɔtto/	*opt* /ˈopt/
‘milk’	*lactem* /ˈlaktẽ/	*leite* /ˈlɐjtə/	*leche* /ˈletʃe/	*latte* /ˈlatte/	*lapte* /ˈlapte/
‘fact’	*factum* /ˈfaktũ/	*feito* /ˈfɐjtu/	*hecho* /ˈetʃo/	*fatto* /ˈfatto/	*fapt* /ˈfapt/

Source: Eifring & Theil (2005)



Source: Wikipedia

Genetically related languages are also geographically contiguous

Languages are also related due to contact over a long period of time

Consequences of language contact

- **Borrowing of vocabulary → loanwords**
- **Adoption of features from other languages**
- Stratal influence
- Language shift

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

Borrowing phonetic form vs semantic content

	form	content	example
Direct loan	Yes	Yes	Avatar, Guru (English) < Sanskrit/Hindi Music (English) < musique (French)
Loanblend	Partly	Yes	double kamrA (Hindi) < double room (Eng) rajasva bajaTa (Hindi) < revenue budget (English)
Loan translation	No	Yes	rajasva ghaTA (Hindi) < revenue budget (English)
Loan creation	No	Yes	prashikshaNArthi (hindi) < trainee (English)
Loanshift	No	Yes	Vidyut (org. lightning) < electricity (English)

Adoption of Features from other languages

- Over a long period of sustained exchange, languages can come closer
- Creation of a *Linguistic Area*
- **Linguistic Area:** A group of languages (at least 3) that have common structural features due to geographical proximity and language contact *(Thomason 2000)*

India	Balkans
Standard Average European	South East Asia

An example: India *(Emeneau, 1956; Subbarao, 2012; Abbi, 2012)*

- Retroflex sounds: Not found in Indo-European outside Indo-Aryan family
- Vocabulary exchanges: IA → Dravidian as well as Dravidian → IA
- Echo words
  - Generally meaning *etc* or *things like this*
  - Hindi*: cAya-vAya (cAya → tea)*
  - *Telugu: puli-guli (puli → tiger)*

and many more: Dative Subjects, Compound & Conjunct Verbs, etc.

To the layperson, Dravidian & Indo-Aryan languages would seem closer to each other than English & Indo-Aryan

What does language relatedness imply for MT?

- 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
  - **hin:** दाल में कुछ काला होना (*dAla me.n kuCha kAlA honA)*
  - **guj:** दाळ मा काईक काळु होवु (*dALa mA kAIka kALu hovu)*

Literal meaning: *something black in the lentils*

Idiomatic meaning: *something fishy*

Tutorial Outline

- Introduction & Motivation
- Language Relatedness
- **Translation within related languages**
- Translation from related languages to another language
- Summary

Translation within related languages

Let’s see how we can use the relatedness between languages to improve translation quality

X and Y are related to each other

In this section, we focus on one key characteristic of related languages - Lexical Similarity

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Lexically Similar Languages

*(Many words having similar **form** and **meaning**)*

- *Cognates*

a common etymological origin

roTI (hi)	*roTlA (pa)*	*bread*
phala.m (te)	*pazha.m (ta)*	*fruit*

- *Named Entities*

do not change across languages

- *Loan Words*

borrowed without translation

matsya (sa)	*matsyalu (te)*	*fish*
pazha.m (ta)	*phala (hi)*	*fruit*

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

dAla me.n kuCha kAlA honA	*(hi)*	*Something fishy*
dALa mA kAlka kALu hovu	(gu)	

Let's just call such words 'orthographically similar'

But, be warned of ……

<u>False Friends:</u> Similar spelling ; different meaning

- *Different origin: pAnI (hi) [water] → panI (ml) [fever]*
- *Semantic shift: bala means hair (hi, frequent sense) and baLa means child (mr)*

<u>Short words:</u>

jaLa ← → jAla

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Compare similarity of grapheme sequences

Hindi → अ ◌ं ध ◌ा प न　　　Marathi → आ ◌ं ध ळ ◌े प ण ◌ा

a.mdhApana　　　A.mdhLepNA

OR

Compare similarity of phoneme sequences

ə n dʱ a p ə n　　　a n dʱ ɭ e p ə ɳ a

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
    - Grapheme based metrics
    - Phoneme based metrics
    - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
    - Why adapt?
    - Augmenting Parallel corpus with lexically similar words
    - Use orthographic features for Word Alignment
    - Transliterate lexically similar OOV words
    - A different paradigm – character-level SMT

x: a.mdhApana (Hindi)

y: A.mdhLepNA (Marathi)

$$\boldsymbol{prefix}(\boldsymbol{x},\boldsymbol{y}) = \frac{len(matching\_prefix(x,y))}{\max(len(x), len(y))}$$

$$= \frac{0}{8} = 0$$

$$\boldsymbol{jaccard}(\boldsymbol{x},\boldsymbol{y}) = \frac{|x \cap y|}{|x| + |y| - |x \cap y|}$$

$$= \frac{4}{10} = 0.4$$

$$\boldsymbol{dice}(\boldsymbol{x},\boldsymbol{y}) = \frac{2 \times |x \cap y|}{|x| + |y|}$$

$$= \frac{8}{14} = 0.57$$

$$\boldsymbol{lcsr}(\boldsymbol{x},\boldsymbol{y}) = \frac{len(longest\_common\_subsequence(x,y))}{\max(len(x), len(y))}$$

$$= \frac{3}{8} = 0.375$$

$$\boldsymbol{ned\_b}(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{edit\_distance(x,y))}{\max(len(x), len(y))}$$

$$= 1 - \frac{5}{8} = 0.375$$

Variants:

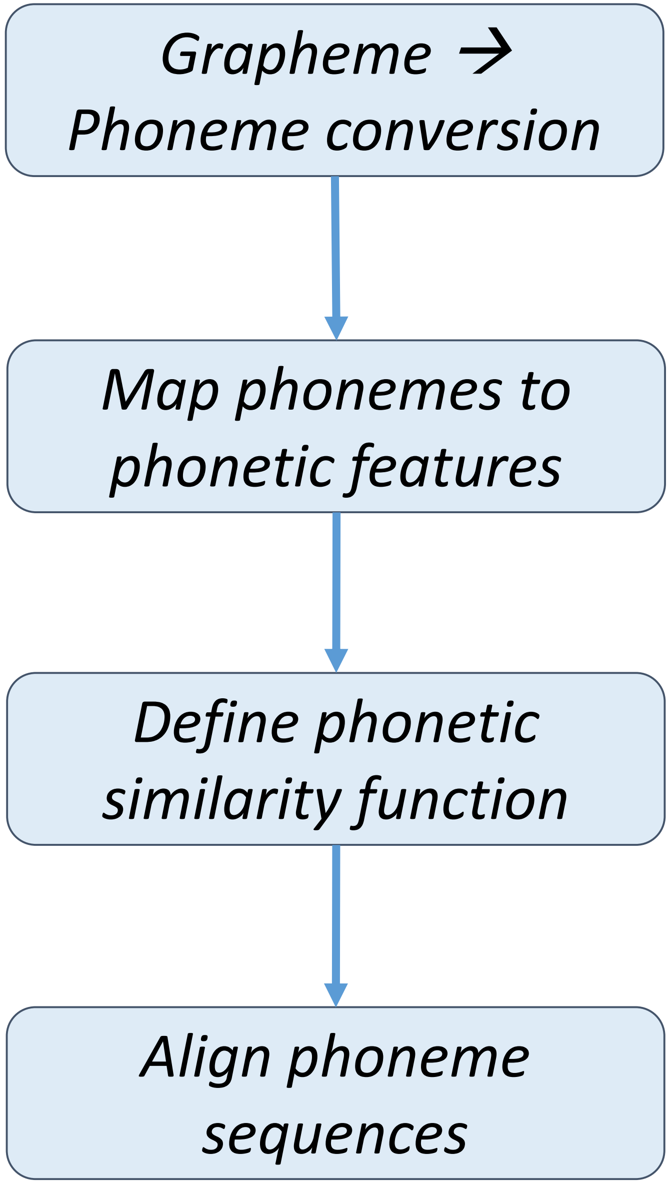
- *Use n-gram as basic unit* *(Inkpen et al,2005)*
- *Skip-gram based metric* *(Inkpen et al,2005)*
- *Similarity matrix to encode character similarity* *(Ristad, 1999; Yarowsky, 2001)*
- LCSF metrics to fix LCSR preference for short *words* *(Kondrak, 2005)*

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT


Grapheme → Phoneme conversion
Map phonemes to phonetic features
Define phonetic similarity function
Align phoneme sequences

x = अ ं ध ा प न → a n dh A p a n

y = आ ं ध ळ े प ण ा → A n dh a L e p a N A

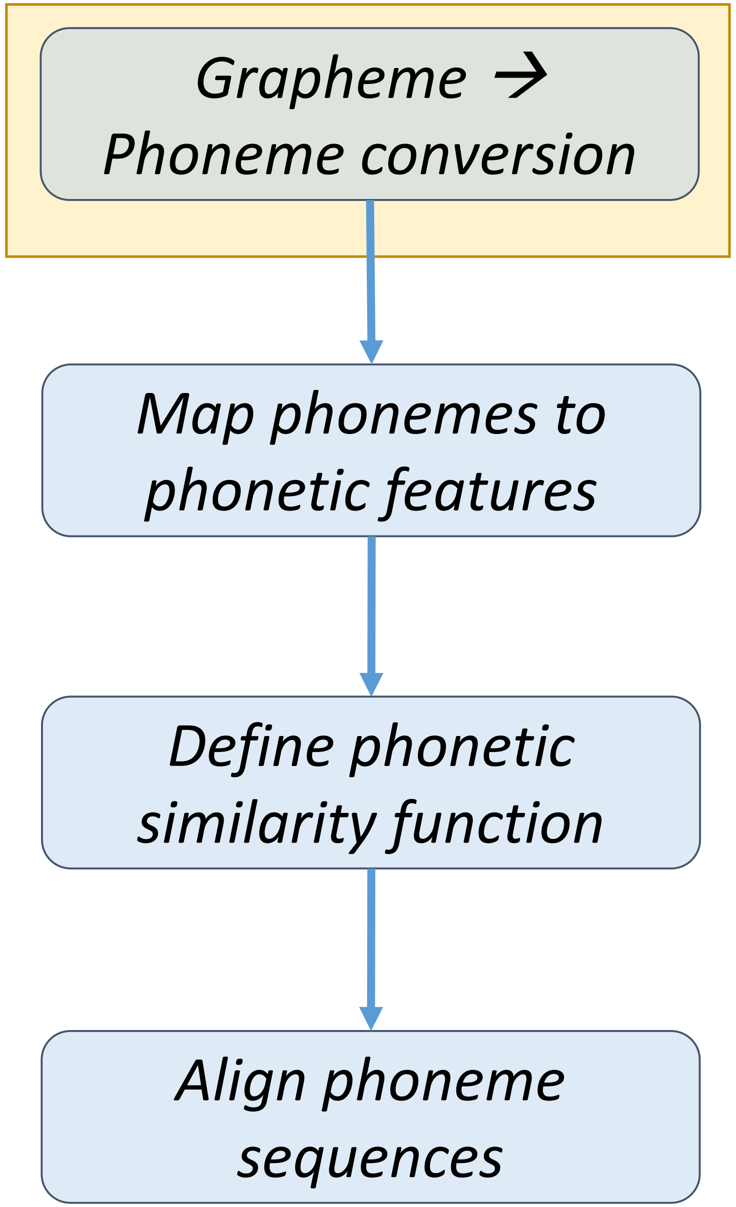
v(‘a’) =(vowel, long , back, open, not_rounded)

v(‘A’) =(vowel, short, back, open, not_rounded)

phonetic_sim(‘a’, ‘A’) = cosine(v(‘a’), v(‘A’))

a n dh A _ e p a n _

A n dh a L e p a N A *sim(x,y)*=6.6



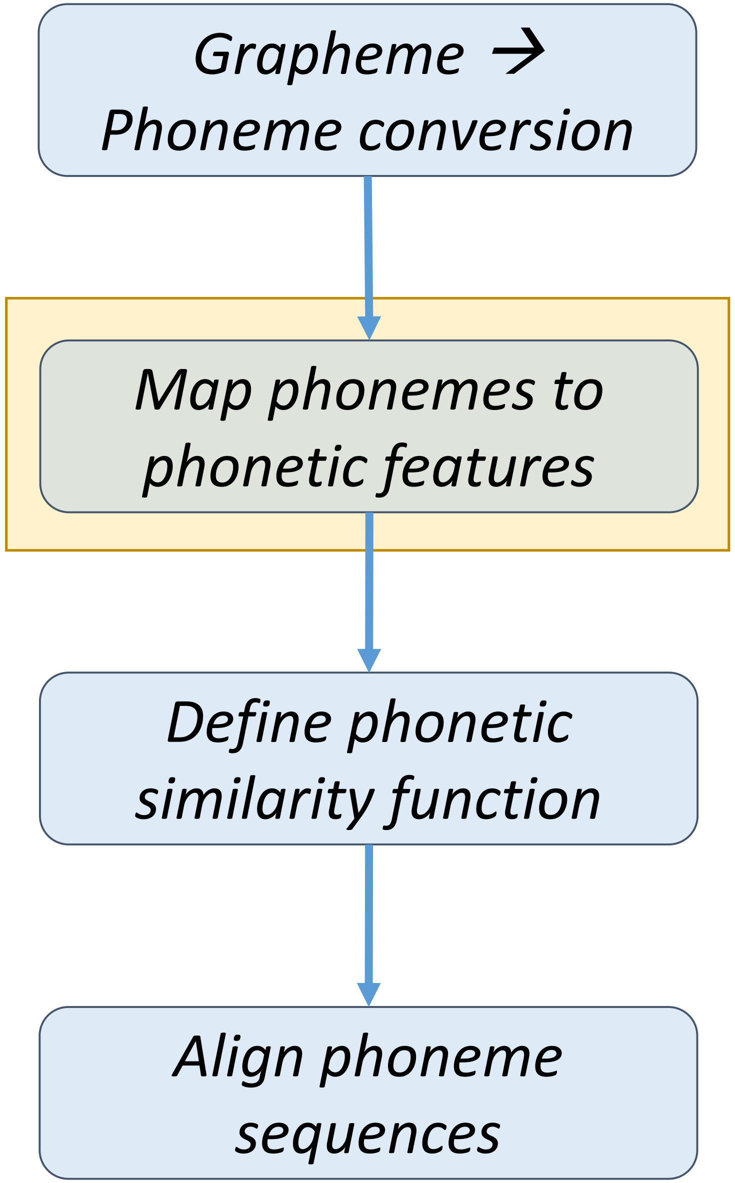
Grapheme → Phoneme conversion

Map phonemes to phonetic features

Define phonetic similarity function

Align phoneme sequences

Some scripts are near phonetic (Brahmi-derived scripts in India) making grapheme → phoneme conversion straightforward



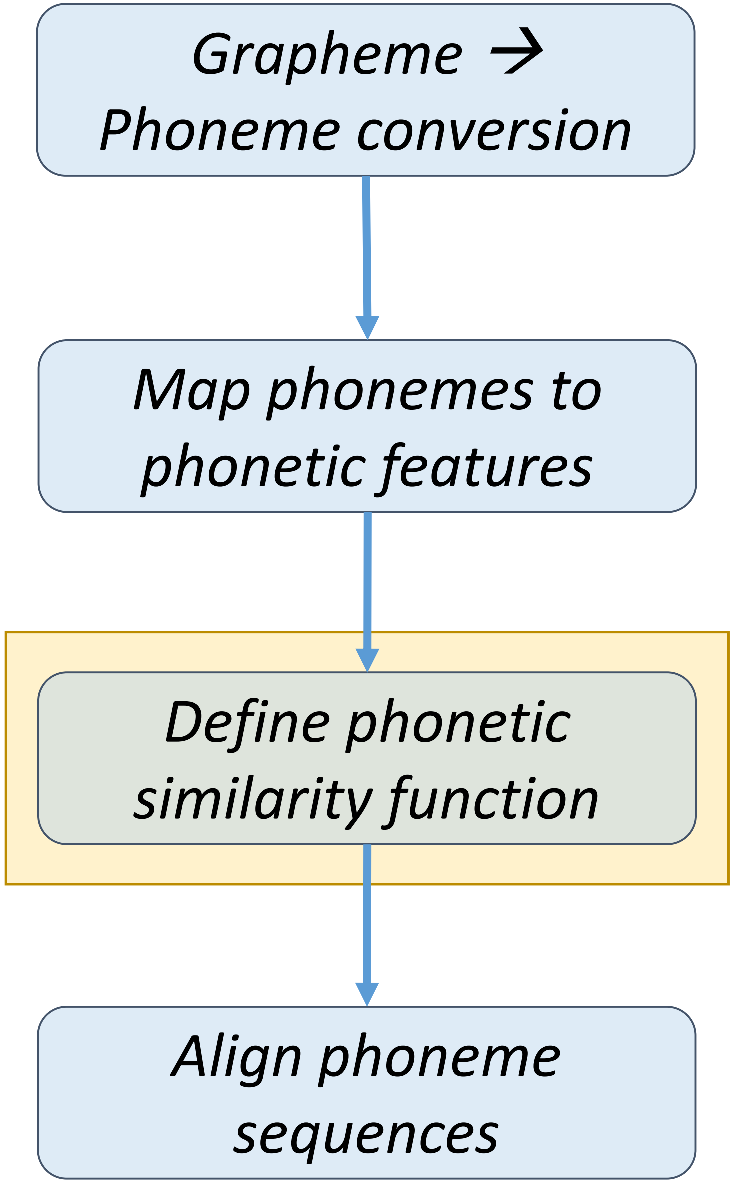
Grapheme → Phoneme conversion

Map phonemes to phonetic features

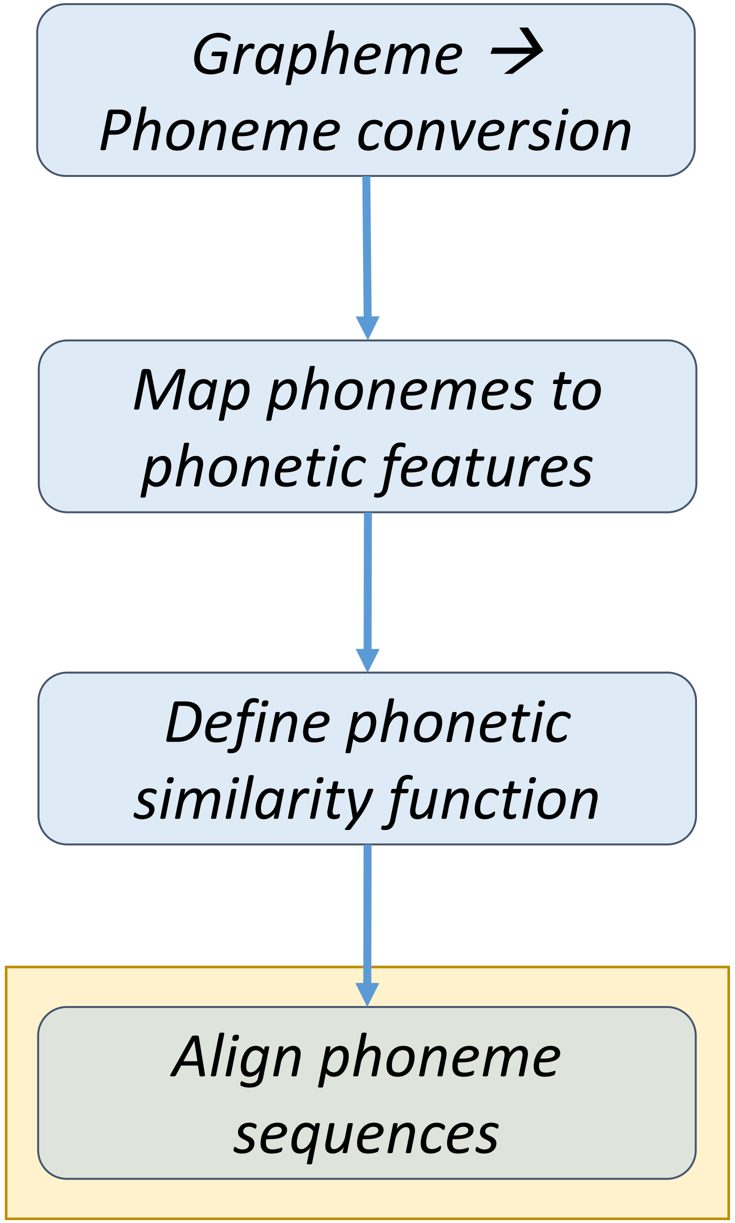
Define phonetic similarity function

Align phoneme sequences

Feature	Values
Basic Character Type	vowel , consonant, nukta, halanta, anusvaara
Vowel Length	short, long
Vowel Strength	weak (a,aa,i,ii,u,uu), medium (e,o), strong (ai,au)
Vowel Status	Independent, Dependent
Vowel – horizontal position	front, back
Vowel – vertical position	open, open-mid, close,close-mid
Vowel – Roundedness	True, False
Consonant Type	plosive, fricative, central approximant, lateral approximant, flap
Place of Articulation	velar,palatal, retroflex, dental, labial
Aspiration	True, False
Voicing	True, False
Nasal	True, False


Grapheme → Phoneme conversion
Map phonemes to phonetic features
Define phonetic similarity function
Align phoneme sequences

Cosine similarity, Hamming, Distance, Handcrafted similarity matrices



Grapheme → Phoneme conversion

Map phonemes to phonetic features

Define phonetic similarity function

Align phoneme sequences

Dynamic Programming, ALINE (Kondrak, 2000)

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - **Putting these metrics to use**

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

- *Thresholding based on similarity metrics*

- *Classification with similarity & other features*
  - *Cognates/False Friends v/s Unrelated*
  - *Cognates v/s False Friends*

- *Competitive Linking*
  - *Similarity based greedy bipartite matching of source words to target cognate candidates*

Cognates/False Friends v/s Unrelated *(Inkpen et al 2005)*

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

Performance of individual measures
Thresholds were learnt using single feature classifier

Classifier	Accuracy
Baseline	63.75
OneRule	95.66
Naïve Bayes	94.84
Decision Trees	95.66
Dec Tree (pruned)	95.66
IBK	93.81
Ada Boost	95.66
Perceptron	95.11
SVM (SMO)	95.46

Results of classification

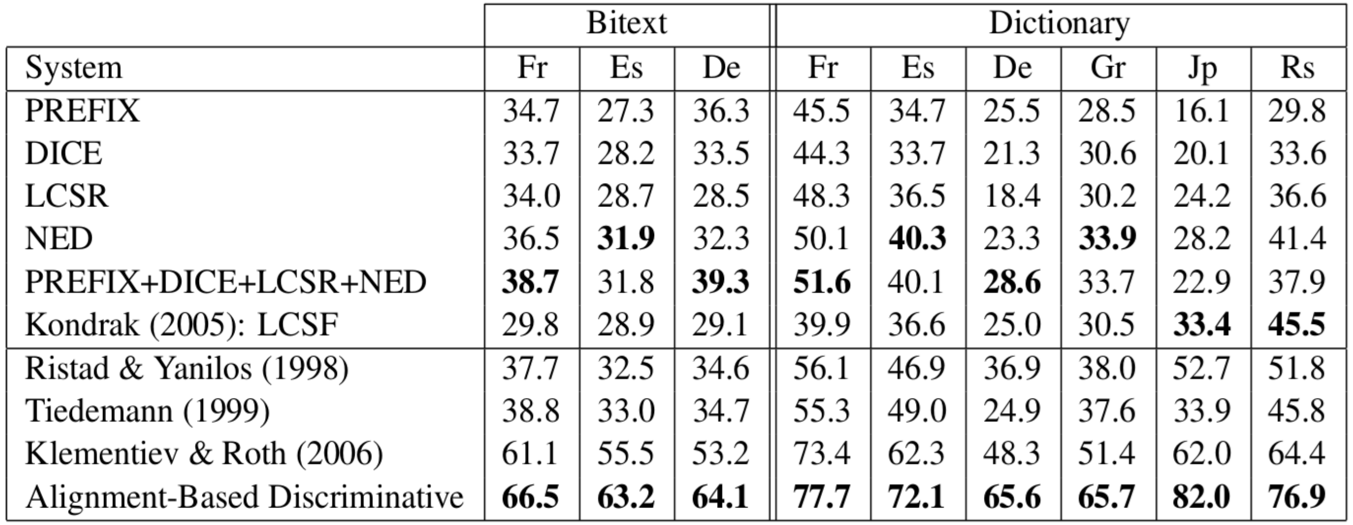
- *LCSR, NED are simple, effective measures*
- *n-gram measures perform well*
- *Classification gives modest improvement over individual measures on this simple task*

Cognates v/s False Friends (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*

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Improve Word Alignment
  - Transliterate lexically similar OOV words

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - **Why adapt?**
  - Augmenting Parallel corpus with lexically similar words
  - Improve Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Limitations of SMT

- No explicit notion of cognates, loanwords and named entities
- All morphological variants of words generally not found in parallel corpus
- Cannot decompose compounds

Consequences

- Sub-optimal word alignment
- Cannot translate unseen cognates and named entities
- Cannot translate morphological variants

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - **Augmenting Parallel corpus with lexically similar words**
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Parallel Corpus

A boy is sitting in the kitchen	Un garçon est assis dans la cuisine
A boy is playing tennis	Un garçon joue au tennis
A boy sitting on a round table	Un garçon assis sur une table ronde
Some men are watching tennis	Certains hommes regardent le tennis
A girl is holding a black book	Une jeune fille tient un livre noir
Two men are watching a movie	Deux hommes regardent un film
abundance	abondance
acrobatic	acrobatique
cabin	cabine
tennis	tennis

How does it help?

Parallel Corpus

A boy is sitting in the kitchen	Un garçon est assis dans la cuisine
A boy is playing tennis	Un garçon joue au tennis
A boy sitting on a round table	Un garçon assis sur une table ronde
Some men are watching tennis	Certains hommes regardent le tennis
A girl is holding a black book	Une jeune fille tient un livre noir
Two men are watching a movie	Deux hommes regardent un film
abundance	abondance
acrobatic	acrobatique
cabin	cabine
tennis	tennis

How does it help?

- *Improves word alignment*
  (10% reduction in word alignment error rate)
- *Improves vocabulary coverage*
- *Improves translation quality*
  (2% improvement in BLEU score as well qualitative improvement)

Parallel Corpus

A boy is sitting in the kitchen	Un garçon est assis dans la cuisine
A boy is playing tennis	Un garçon joue au tennis
A boy sitting on a round table	Un garçon assis sur une table ronde
Some men are watching tennis	Certains hommes regardent le tennis
A girl is holding a black book	Une jeune fille tient un livre noir
Two men are watching a movie	Deux hommes regardent un film
abundance	abondance
acrobatic	acrobatique
cabin	cabine
tennis	tennis

Some tips

- *Focus on high recall in cognate extraction (Kondrak et al, 2003; Onaizan, 1999)*
- *Replication of cognate pairs improves alignment quality marginally (Kondrak et al, 2003; Och & Ney, 1999; Brown et al, 1993)*
- *Add multiple cognate pairs per line (Kondrak et al, 2003)*

pAnI jala nIra ←→ pANI jaLa nIra

Parallel Corpus

A boy is sitting in the kitchen	Un garçon est assis dans la cuisine
A boy is playing tennis	Un garçon joue au tennis
A boy sitting on a round table	Un garçon assis sur une table ronde
Some men are watching tennis	Certains hommes regardent le tennis
A girl is holding a black book	Une jeune fille tient un livre noir
Two men are watching a movie	Deux hommes regardent un film
abundance	abondance
acrobatic	acrobatique
cabin	cabine
tennis	tennis

Limitations

- *Cannot align unseen cognate pairs*
- *Cannot translate unseen words*
- *Knowledge locked in cognate corpus is underutilized*

Lets see if we can overcome some of these limitations pertaining to unseen words

There will still be some unseen words which need to be handled

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - **Use orthographic features for Word Alignment**
  - Transliterate lexically similar OOV words
  - A different paradigm – character-level SMT

Using orthographic features for Word Alignment

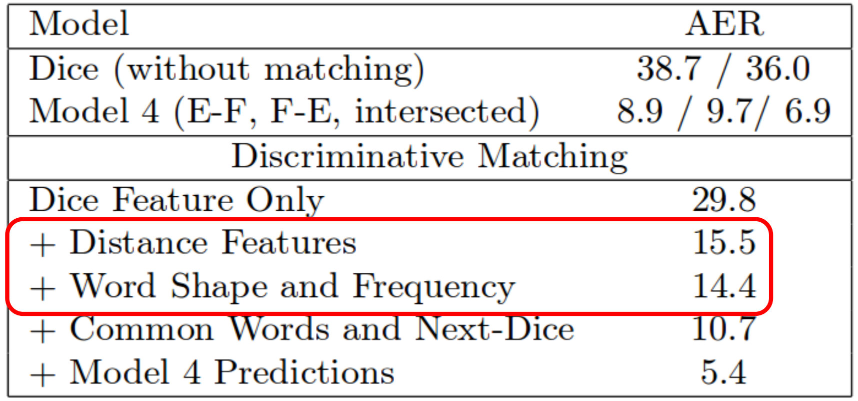
Discriminative models allow incorporation of arbitrary features

(*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*

- *Similar features can be designed for other writing systems*



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)

7% reduction in alignment error rate

There will still be some unseen words which need to be handled

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - **Transliterate lexically similar OOV words**
  - A different paradigm – character-level SMT

Transliterating OOV words

- OOV words can be:
  - **Cognates**
  - **Loan words**
  - **Named entities**
  - Other words
- Cognates, loanwords and named entities are orthographically similar
- *Transliteration achieves translation*
- Orthographic mappings can be learnt from a parallel transliteration/cognate corpus
  - Can be mined from the parallel corpus *(Sajjad et al., 2012; Kunchukuttan et al, 2015)*

Transliterating OOV words

- *Two options*
  - Transliteration as a post-translation step
  - Integrating transliteration into the decoder

Transliteration as Post-translation step

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

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

But first transliteration may not be correct!

Option 2: Generate top-k candidates for each OOV. Each regenerated candidate sentence 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 transliterations

Integrate Transliteration into the Decoder

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

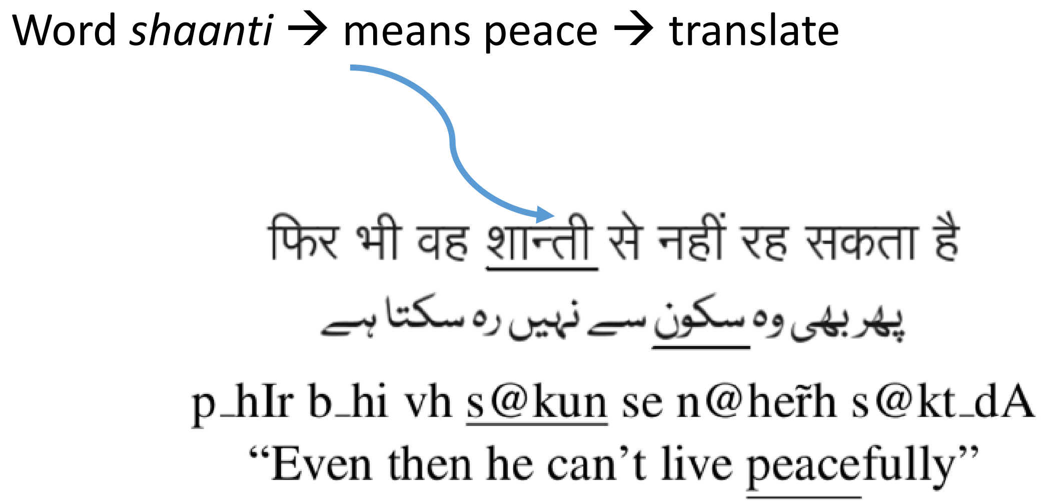
- In addition to translation candidates, decoder considers all transliteration candidates for each word
  - Assumption: 1-1 correspondence between words in the two languages
  - monotonic decoding
- Translation and Transliteration candidates <u>compete</u> with each other
- The features used by the decoder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

Results (Hindi-Urdu Translation)

Durrani et al (2010)

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

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



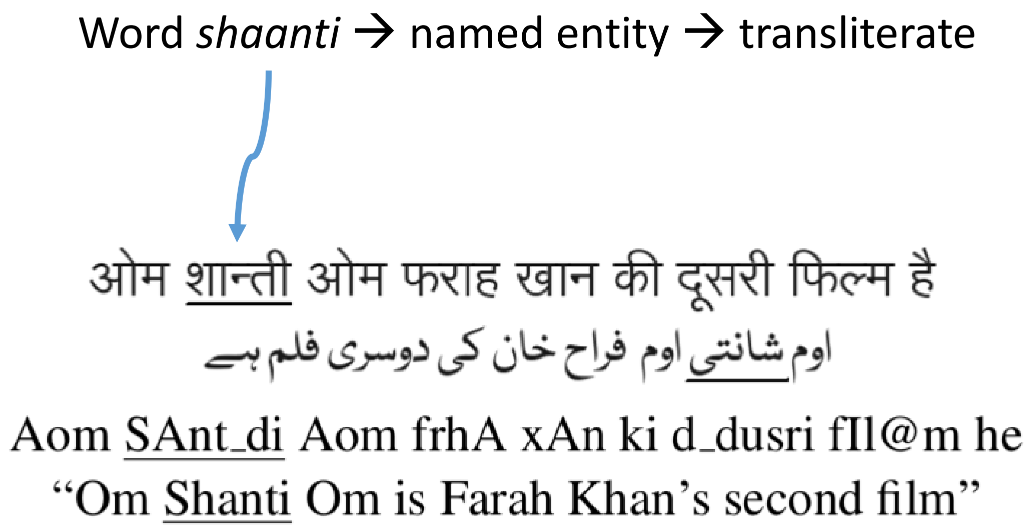
Word *shaanti* → means peace → translate

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@her̃h s@kt_dA

“Even then he can’t live peacefully”



Word *shaanti* → named entity → transliterate

ओम शान्ती ओम फराह खान की दूसरी फिल्म है

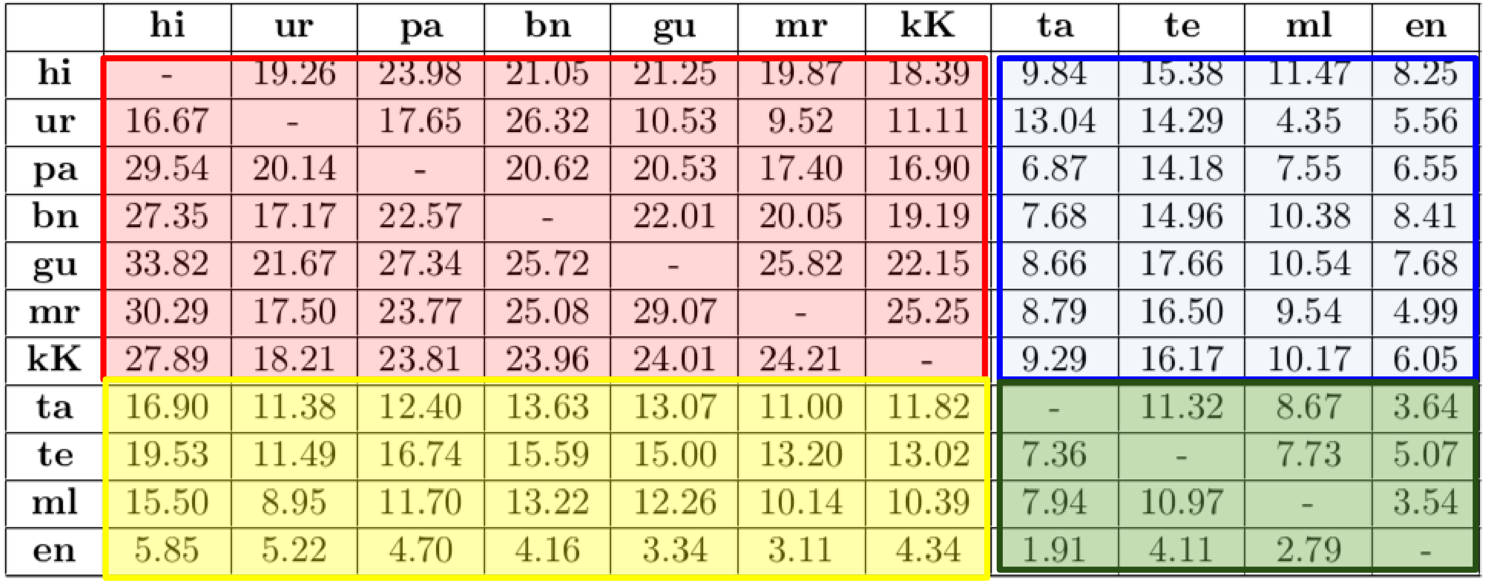
اوم شانتی اوم فراح خان کی دوسری فلم ہے

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)



		Indo-Aryan							Dravidian			
		hi	**ur**	**pa**	**bn**	**gu**	**mr**	**kK**	**ta**	**te**	**ml**	**en**
Indo-Aryan	**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
Dravidian	**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	-

% OOV decrease after transliterating untranslated words

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

The story so far….

Leverage Lexical Similarity by Adapting Word Level SMT…

So far so good….

But there are some shortcomings…

Shortcomings of Adapting word-based methods

- Additional resources and tools required
  - Cognate corpus
  - Transliteration corpus
  - Word aligned corpus
  - Morphological analyzers

- Not directly optimized for improving SMT performance

We are "retrofitting" a word-level system to incorporate lexical similarity

Is word the right level of representation for translation?

Explore sub-word units of representation for translation

Roadmap for this section

- **What** is Lexical Similarity?

- **How** to identify lexically similar words?
  - Grapheme based metrics
  - Phoneme based metrics
  - Putting these metrics to use

- **Why** focus on lexical similarity?
  (Or Adapting SMT for leveraging lexical similarity)
  - Why adapt?
  - Augmenting Parallel corpus with lexically similar words
  - Use orthographic features for Word Alignment
  - Transliterate lexically similar OOV words
  - **A different paradigm – character-level SMT**

Basic unit of translation → CHARACTER

Transliteration for translation

	Word-level	**Character-level (unigram characters)**
hi	राम ने श्याम को पुस्तक दी	र ा म _ न े _ श ् य ा म _ क ो _ प ु स ् त क _ द ी
	rAma ne shyAma ko pustaka dI	r A ma _ n e _ sh y A ma _ k o _ p u s ta ka _ d I
mr	रामाने श्यामला पुस्तक दिली	र ा म ा न े _ श ् य ा म ल ा _ प ु स ् त क _ द ि ल ी
	rAmAne shyAmalA pustaka dilI	r A m A n e _ sh y A ma l A _ p u s ta ka _ d i l I

Gloss	Ram+*nom* Shyam+*acc* book gave
English Translation	Ram gave a/the book to Shyam

Why character-level SMT?

High degree of character-level similarity between related languages

LCSR as a measure of language relatedness
(computed at sentence level on a parallel corpus)

Konkani – Marathi	54.51
Punjabi – Hindi	68.00
Bulgarian – Macedonian	62.85
Danish – Swedish	63.39
Indonesian – Malay	73.54

Primary language divergences can be bridged by sub-word transformations

- Spelling/pronunciation differences (Cognates, Loan words)
- Suffix sets & function words: mappings can be learnt for short sequences

cA → kA madhye → me.m (for Marathi → Hindi)

An integrated framework tackling cognates, named entities, inflection, agglutination

Training Character level SMT

Use the same discriminative log-linear framework as Phrase-based SMT

… with some modifications …

Modification 1: Handling sentence length issues during training

Long sentences at character level → Inefficient Word alignment

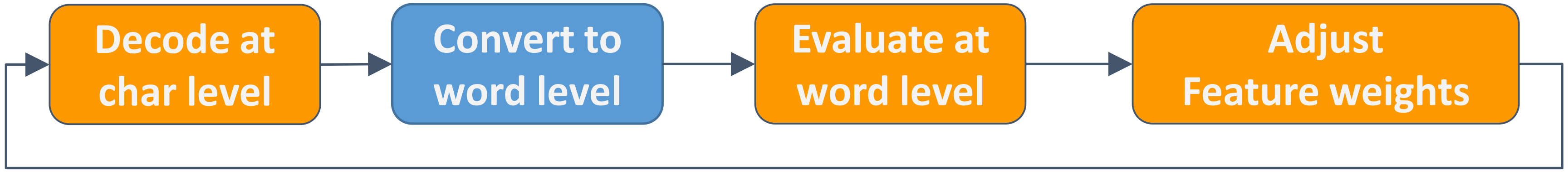
(a) Limit sentence length → Loss of training corpus *(Tiedemann, 2009)*
(b) Phrase pairs from word-based model as corpus → Larger models *(Vilar, 2007)*

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

Modification 2: Monotone decoding (Tiedemann, 2009)

Modification 3: Tuning at word-level (Tiedemann, 2012)

MERT Tuning



Further improvements to character-based SMT …

(Tiedemann, 2009; Nakov & Tiedemann, 2012; Tiedemann & Nakov, 2013)

- Longer units of translation: character n-grams
  - n>2 has not been useful
- Capturing larger context information → higher order LM and longer phrase-pairs
  - Data sparsity a lesser issue
  - Improves translation quality
- Combining word and character models useful
  - System combination
  - Merging phrase tables
- Filtering noisy entries in phrase tables improves quality

Can suffixes & function words be translated?

Function words (which differ across related languages) can be learnt

kok: ह्या किड्याचें खाशेलपण कळ्ळे उपरांत दिसता तांचो संवसारूय कितलो मजेशीर आसा .

hyA kiDyAce.n khAshelapaNa kaLLe uparA.nta disatA tA.nco sa.nvasAruuya kitalo majeshIra AsA

mar: ह्या किड्याची विशेषता कळल्यानंतर दिसते त्यांचे विश्वस्देखील किती मजेदार आहे .

hyA kiDyAcI visheShatA kaLalyAna.ntara disate tyA.nce vishvaradekhIla kitI majedAra Ahe .

gloss: these insects_of uniqueness knowing_after see their world_also how funny is

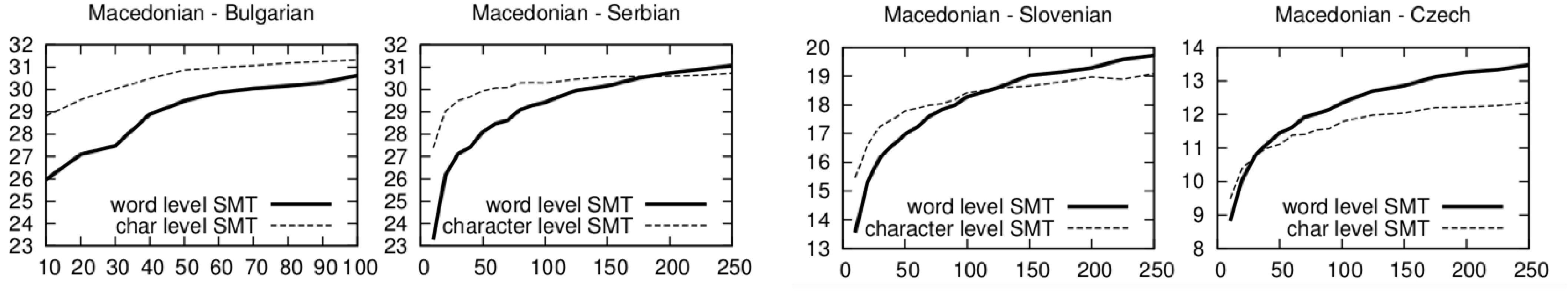
eng: After knowing the uniqueness of these insects, <we> realize how funny their world is.

Even content words which are not orthographically similar can be learnt

Is character-level SMT good for small corpora?

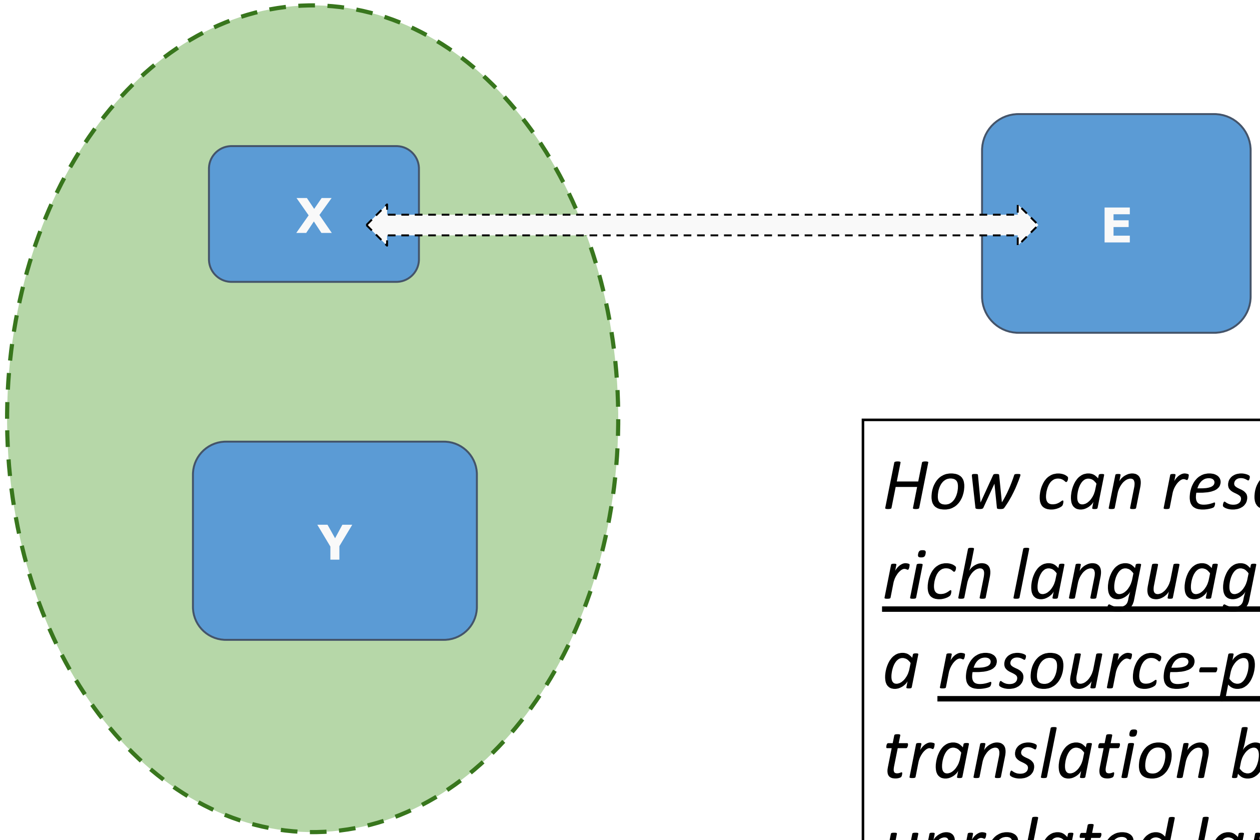
- Character-level SMT can outperform word-level when very little corpus is available
- With increased parallel corpus, the performance gap narrows
- *<u>The similarity between the source and target languages is also important</u>*

  (Czech is not as close to Macedonian as others)



Tutorial Outline

- Introduction & Motivation
- Language Relatedness
- Translation within related languages
- **Translation from related languages to another language**
- Summary

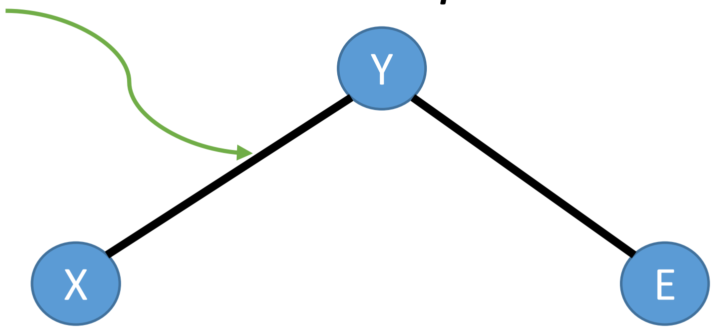


How can resources for a <u>resource-rich language</u> Y, which is related to a <u>resource-poor language</u> X, help translation between X, and an unrelated language E?

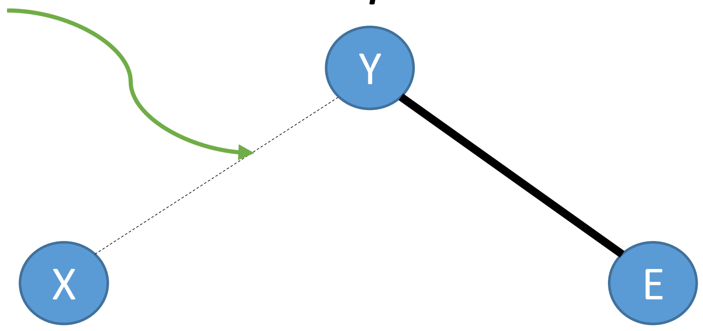
Scenarios based on corpus availability….

Y: bridge/pivot language

Sufficient Parallel Corpus



No or little Parallel Corpus



- *Scenario can occur between unrelated languages too*
- *Does not necessarily leverage relatedness between languages*

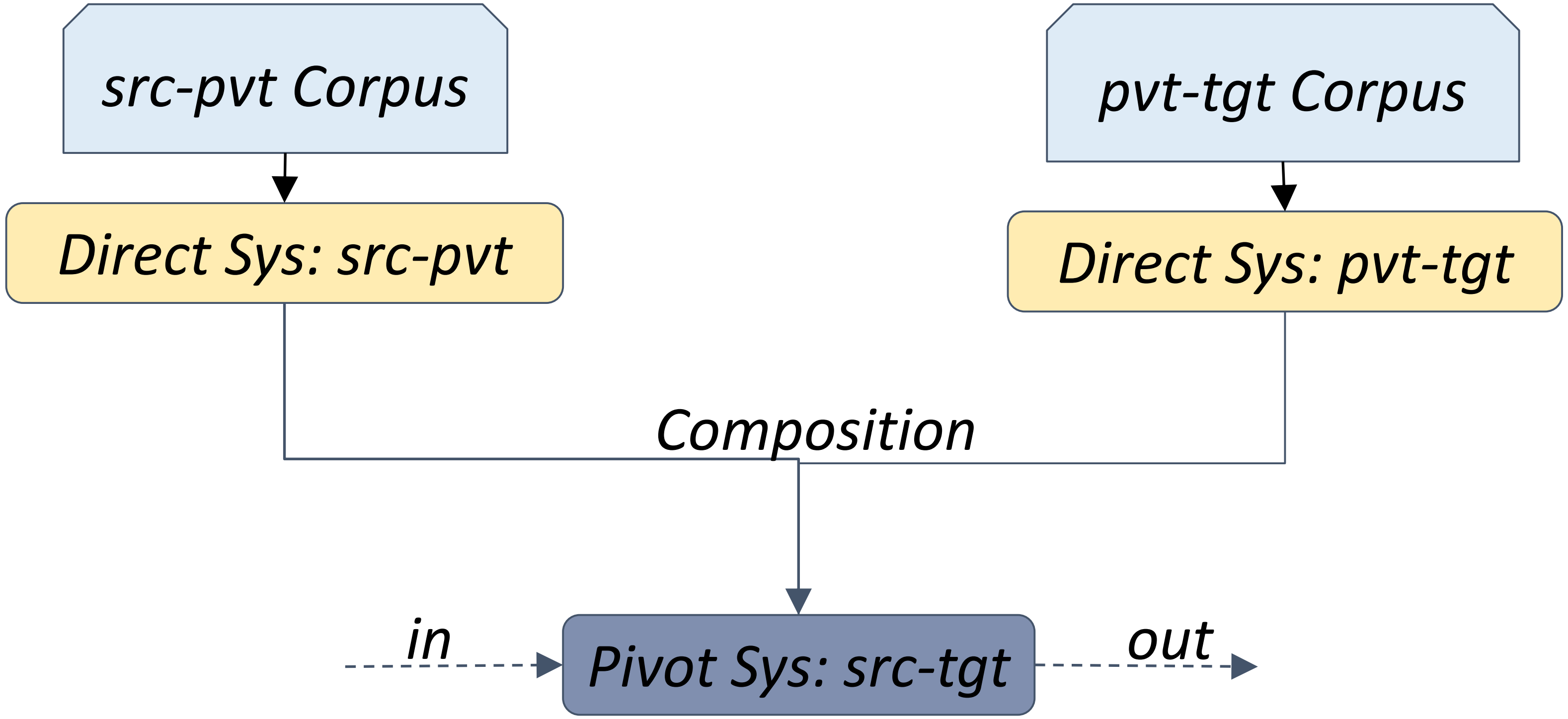
- *Relatedness between X and Y will have to be leveraged*

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Roadmap for this section

- **Pivot based SMT**
    - Pseudo-Corpus Synthesis
    - Cascading Direct Systems
    - Model Triangulation
    - Case Study I
- **Leveraging relatedness in Pivot based SM**
    - Small X→Y corpus is available (Case Study II)
    - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
    - Combine corpus
    - Combine models
- **Choice of pivot language**

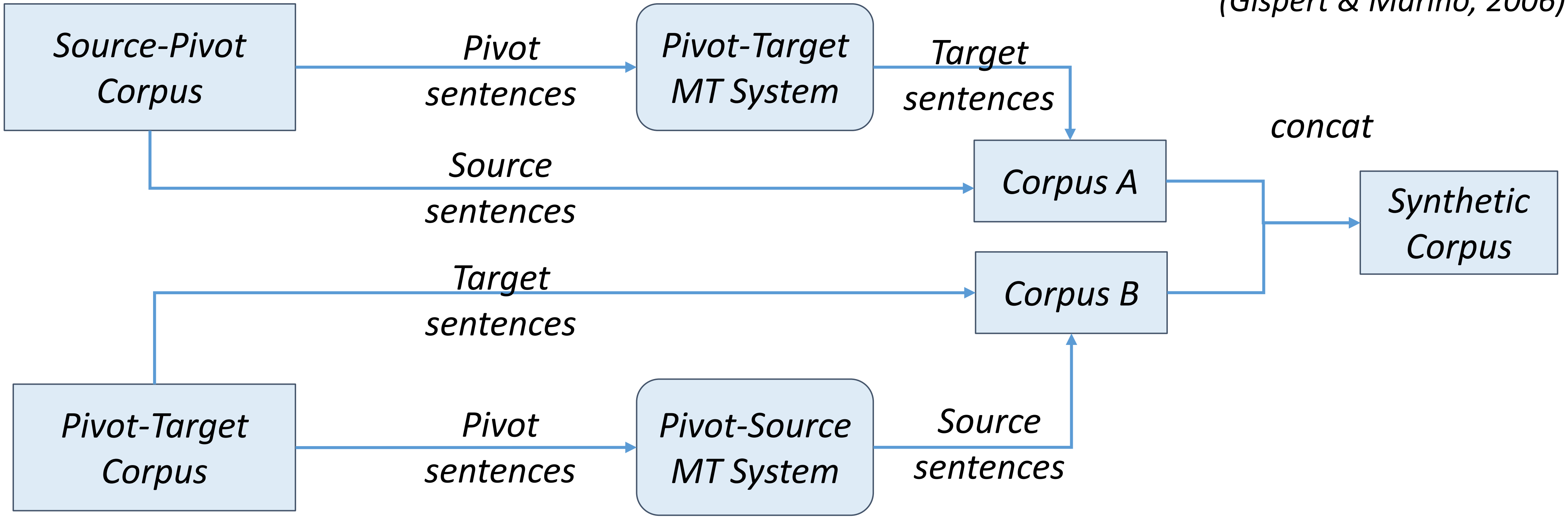

src-pvt Corpus
pvt-tgt Corpus
Direct Sys: src-pvt
Direct Sys: pvt-tgt
Composition
in
Pivot Sys: src-tgt
out

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Goal: Create a Pseudo Source-Target training corpus

(Gispert & Marino, 2006)



Generated corpus will be noisy; quality would depend on:

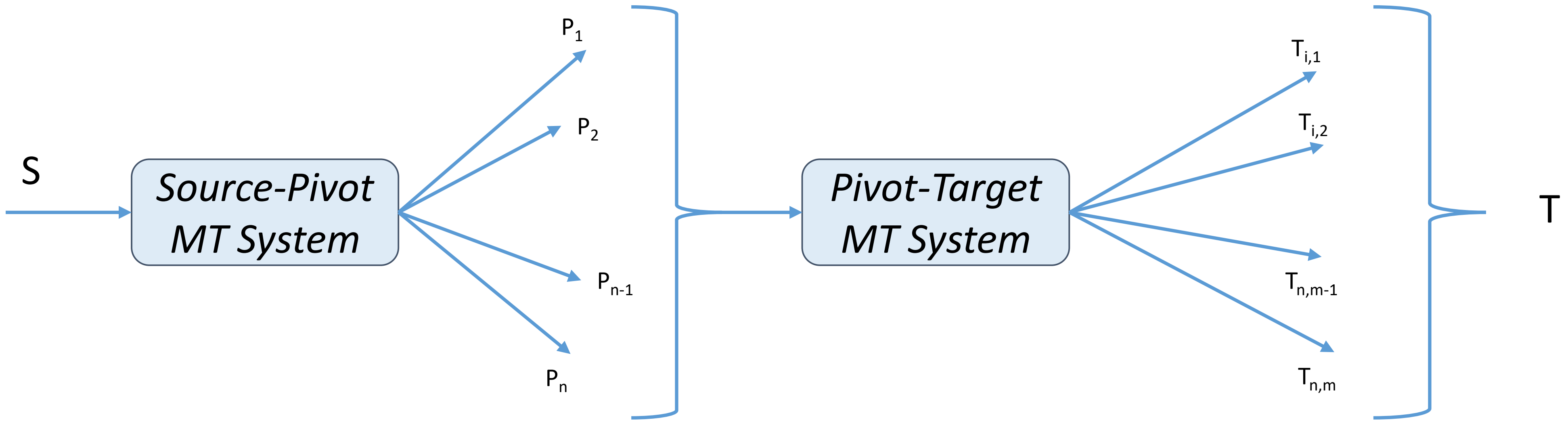
(i) language divergence (ii) parallel corpus size

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

(Utiyama & Isahara, 2007)

Top-n candidates



Re-rank the m.n target language candidates by interpolating scores

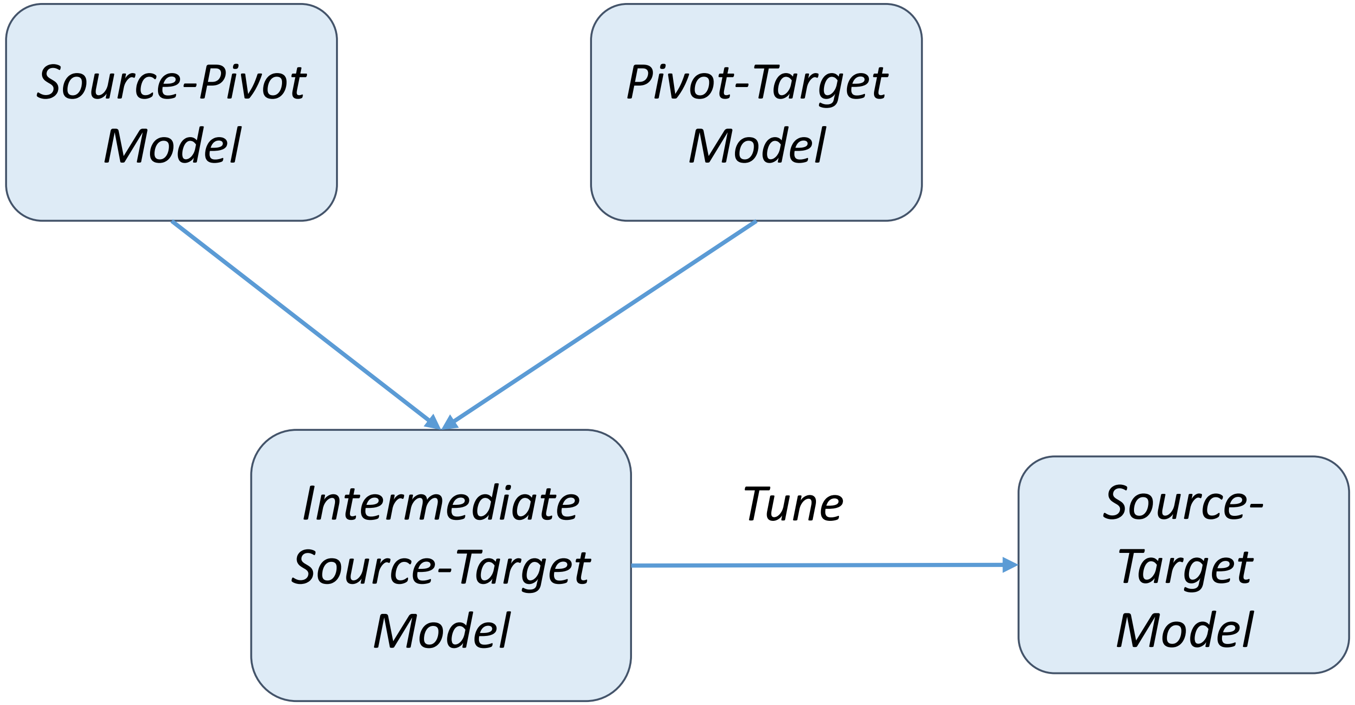
$$t = \operatorname*{argmax}_{t \in T} \sum_{k=1}^{L} \left(\lambda_k^{sp} h_k^{sp}(s,p) + \lambda_k^{pt} h_k^{pt}(p,t) \right)$$

(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

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

(Utiyama & Isahara, 2007; Wu & Wang, 2007)



src-pivot phrase table

A	X	0.4	0.4
B	X	0.6	0.8
B	Y	0.8	0.9
C	Y	0.2	0.1

X	P	0.5	0.4
Y	P	0.5	0.6
Y	Q	1.0	1.0
Z	R	1.0	1.0

pivot-tgt phrase table

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

Comparison

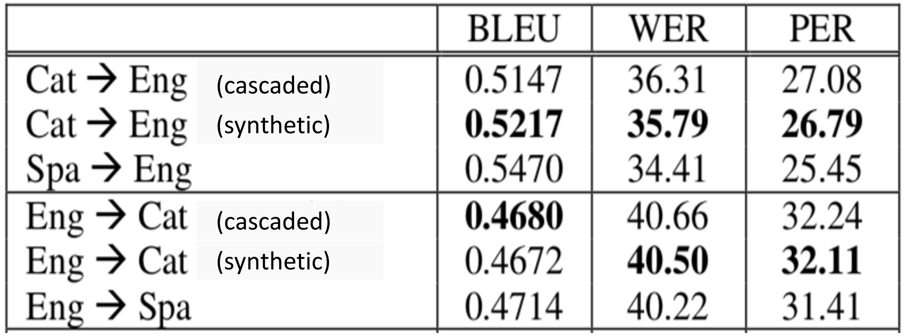
Criteria	Pseudo-corpus	Cascaded	Triangulation
Ease of implementation	Easy	Easy	Involved
Training Time	Depends on time to decode time to created pseudo-parallel corpus	No separate training	High, due to the time required for merging
Decoding Time	Low, just as much as a baseline PBSMT system	Very high, due to multiple decoding	High due to increase in model size
Model Size	same order as PBSMT model of this size training corpus size <=2*max(src-pvt,pvt-tgt) corpus	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

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Case Study I

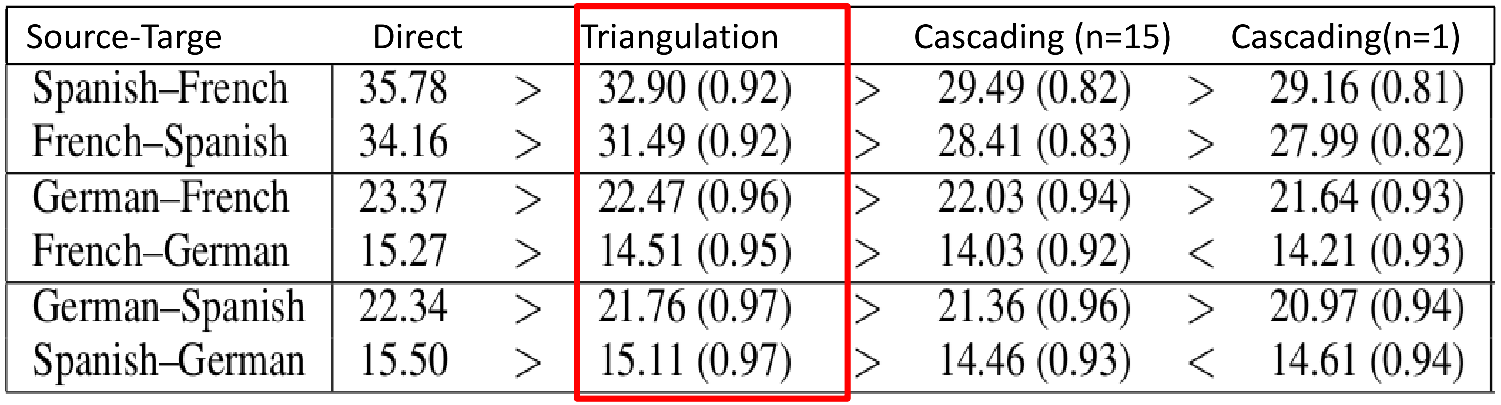
Catalan-English with Spanish as pivot



	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

Marino & Gispert, 2006

English as Pivot



Source-Targe	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)

Utiyama & Isahara, 2007

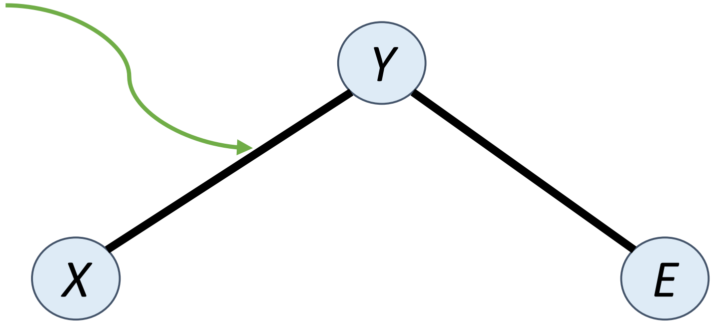
Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Scenarios based on corpus availability….

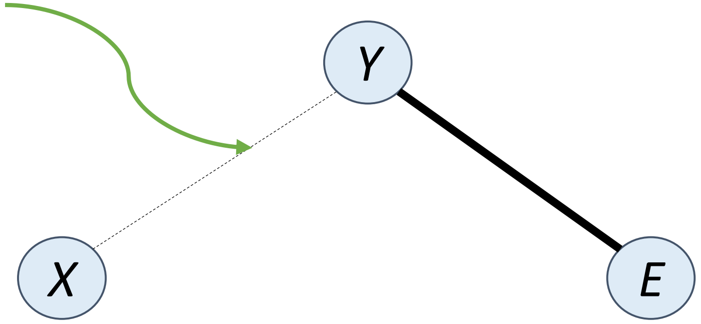
Y: bridge/pivot language

Sufficient Parallel Corpus



- *Scenario can occur between unrelated languages too*
- *Does not necessarily leverage relatedness between languages*

No or little Parallel Corpus

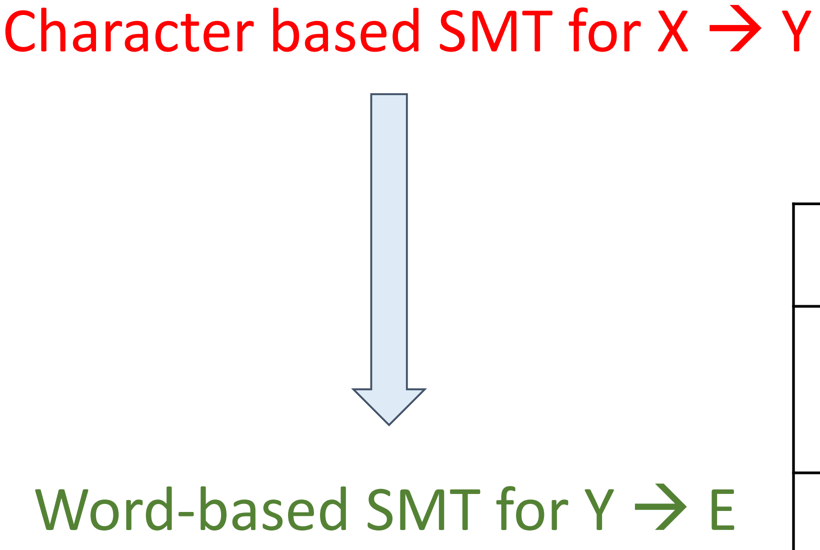


- *Relatedness between X and Y will have to be leveraged*

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Case Study II *(Tiedemann, 2012)*



Character based SMT for X → Y

Word-based SMT for Y → E

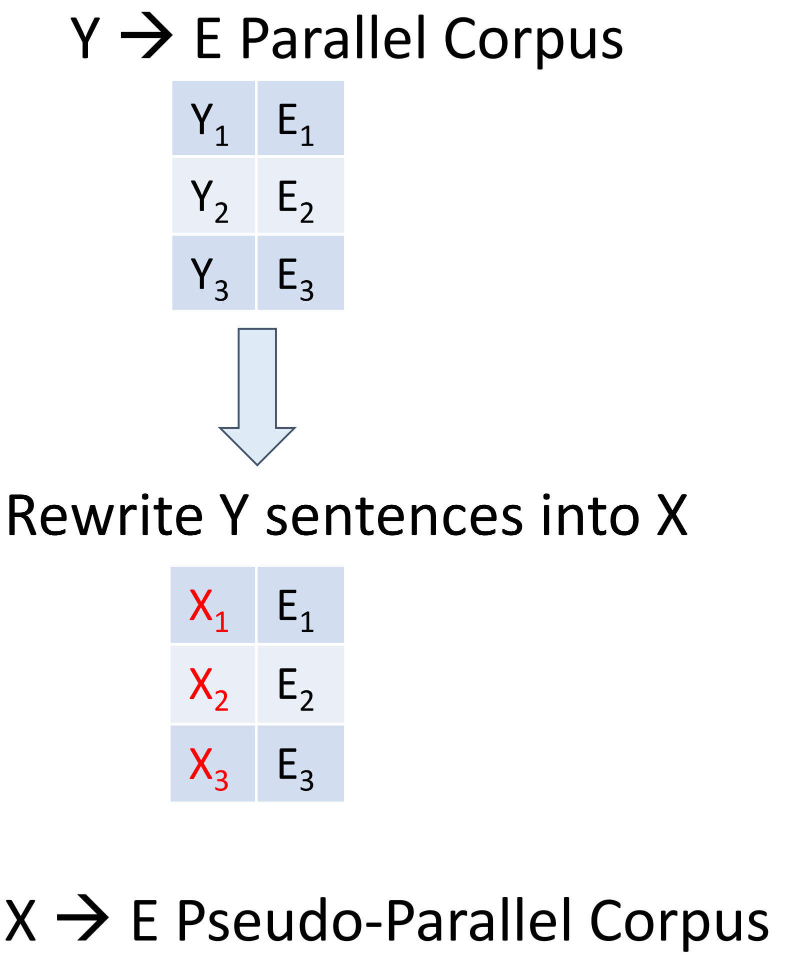
- Char-based SMT effective with small corpora
- X → Y leg of pivot SMT may generate non-words

			X→ E (% BLEU)			X→ Y (% BLEU)		OOV % char level
X	**E**	**Y**	**Direct**	**Pivot-word**	**Pivot-char**	**word-level**	**char-level**	
mk	en	bs	20.74	12.48	18.64	14.22	**24.82**	1.00
		bg		19.74	**21.10**	14.77	**17.28**	0.77
gl	en	es	*5.76*	13.2	**16.02**	43.22	**50.70**	1.36
ca	en	es	27.86	38.65	**40.73**	59.34	**65.14**	0.48

- Macedonian (X) is related to Bulgarian (Y) and Bosnian (Y)
- Galician (X) and Catalan (X) are related to resource rich Spanish (Y)
- X-Y corpus in thousands, while Y-E (English) corpus in millions

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**



Y → E Parallel Corpus

Y_1	E_1
Y_2	E_2
Y_3	E_3

Rewrite Y sentences into X

X_1	E_1
X_2	E_2
X_3	E_3

X → E Pseudo-Parallel Corpus

For each word in Y

- No knowledge sources
  - Do Nothing: Pretend Y is X
- Transliteration or cognate pairs between Y and X
  - Transliterate Y into X
- Word and/or Phrase dictionary between Y and X
- Parallel corpus with a third language Z
  - Induce a word and/or phrase dictionary by pivoting via a third language
- Morphological analyzer for Y and X
  - Generate morphological variants of X from stems in Y

Case Study III *(Wang 2012)*

X: Indonesian Bahasa, Y: Malay, E: English

System	**BLEU %**
Direct X → E (baseline)	18.67
Pretend Y is X	14.50
Rewriting of Y → X	
CN: word dictionary from pivot	19.50
(A) CN: word dictionary from pivot + morph	20.06
(B) CN: phrase dictionary from pivot + morph	20.89
System Combination (A) + (B)	21.24
Adaptation of X → Y (decode time)	
CN: word dictionary from pivot	17.22

- Source rewriting performs better than system trained on a small X → E parallel corpus
- Rewriting of X→Y does not perform
    - Done at decode time
    - Training corpus more robust to noise

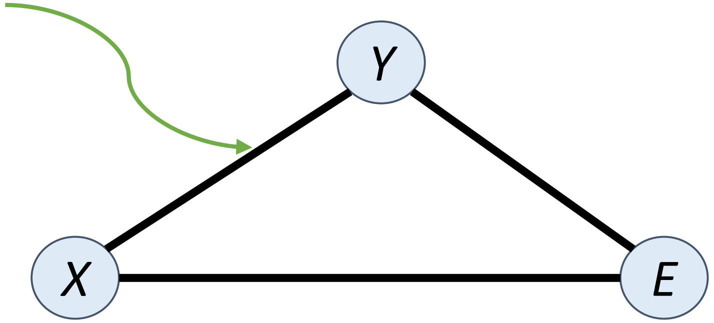
Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

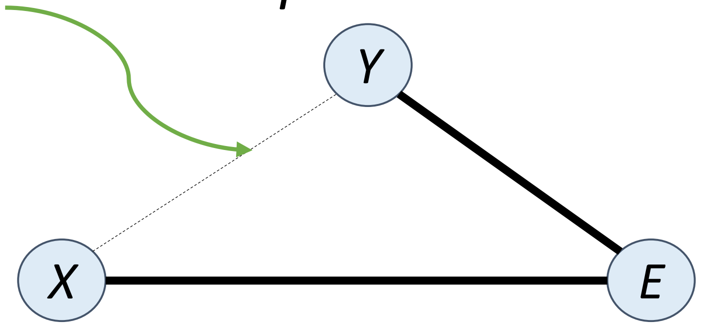
Now suppose we have a parallel corpus between X and E as well

Y: bridge/pivot language

Sufficient Parallel Corpus



No or little Parallel Corpus

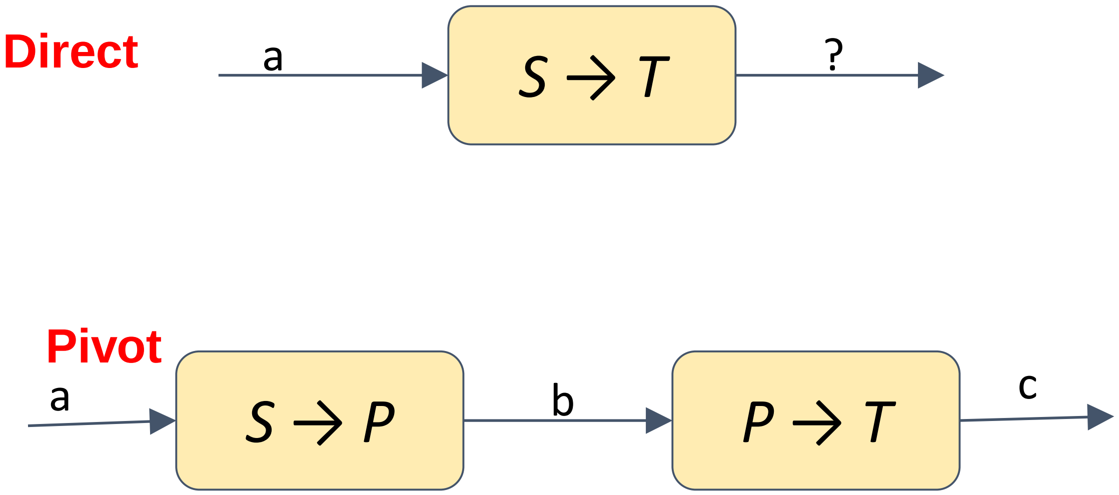


How do we augment direct system with the pivot system?

Can the pivot system improve the direct system?

Improve lexical coverage

Unknown words



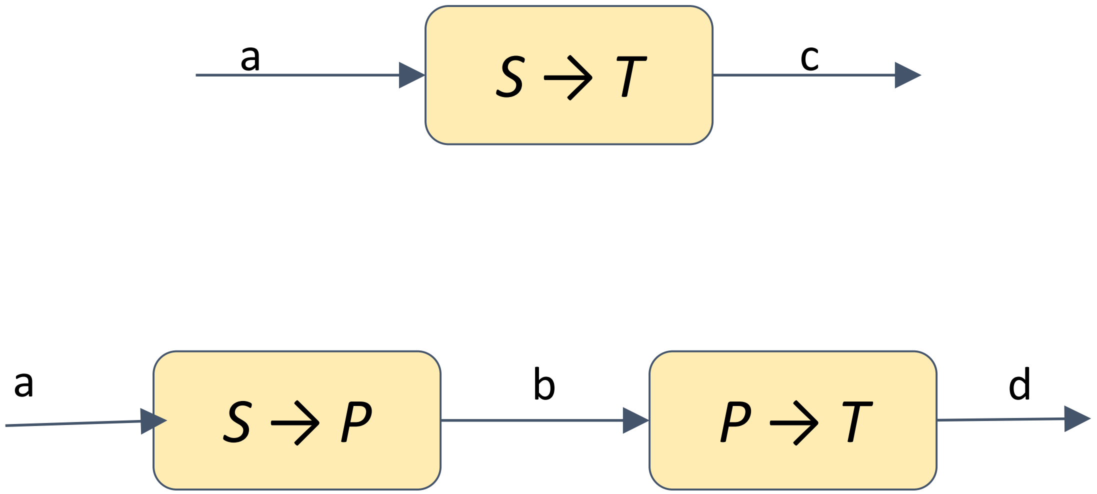
Direct

a → S → T → ?

Pivot

a → S → P → b → P → T → c

More translation options



a → S → T → c

a → S → P → b → P → T → d

Improve Probability estimates

by combining feature values from both tables

Such combination may be useful for translation between related languages too

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Case Study IV *(Wang et al., 2012)*

X: Indonesian Bahasa, Y: Malay, E: English

Adaptation Method	Simple Concat	Balanced Concat	Sophisticated Comb.
Pretend Y is X	18.49	19.79	**20.10**
CN: word dictionary from pivot + morph	20.60	**21.15**	21.05
CN: word dictionary from pivot + morph	21.01	**21.31**	20.98
System Combination	21.55	**21.64**	21.62

Is concatenating corpora better than pivoting in this scenario?
Nakov & Tiedemann, 2009 experiment when no adaptation is done:

- Simple concatenation cannot be shown to be better
- Sophisticated concatenation is better
- No study for the case of adaptation

Roadmap for this section

- **Pivot based SMT**
  - Pseudo-Corpus Synthesis
  - Cascading Direct Systems
  - Model Triangulation
  - Case Study I
- **Leveraging relatedness in Pivot based SM**
  - Small X→Y corpus is available (Case Study II)
  - No X→Y corpus is available (Case Study III)
- **Augmenting Direct system with Pivot Based System**
  - Combine corpus
  - Combine models
- **Choice of pivot language**

Model 1: Direct model
Model 2: Pivot based model

Combining Model 1 & 2

- Fillup interpolation - Create a unified phrase table – start filling entries from models in order of priority (Dabre et al, 2015)

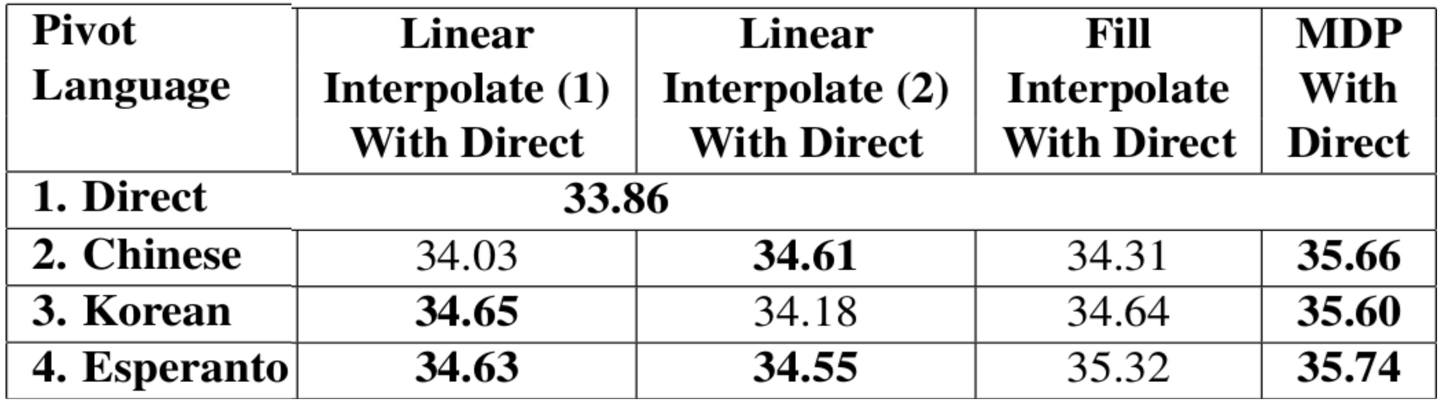
- Linear interpolation – Weighted combination of models (Wu & Wang,2009)

- Multiple decoding paths – Decoder searches over all phrase tables (Nakov & Ng, 2009 ; Dabre et al, 2015)

Case Study V *(Dabre et al., 2015)*

- 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**

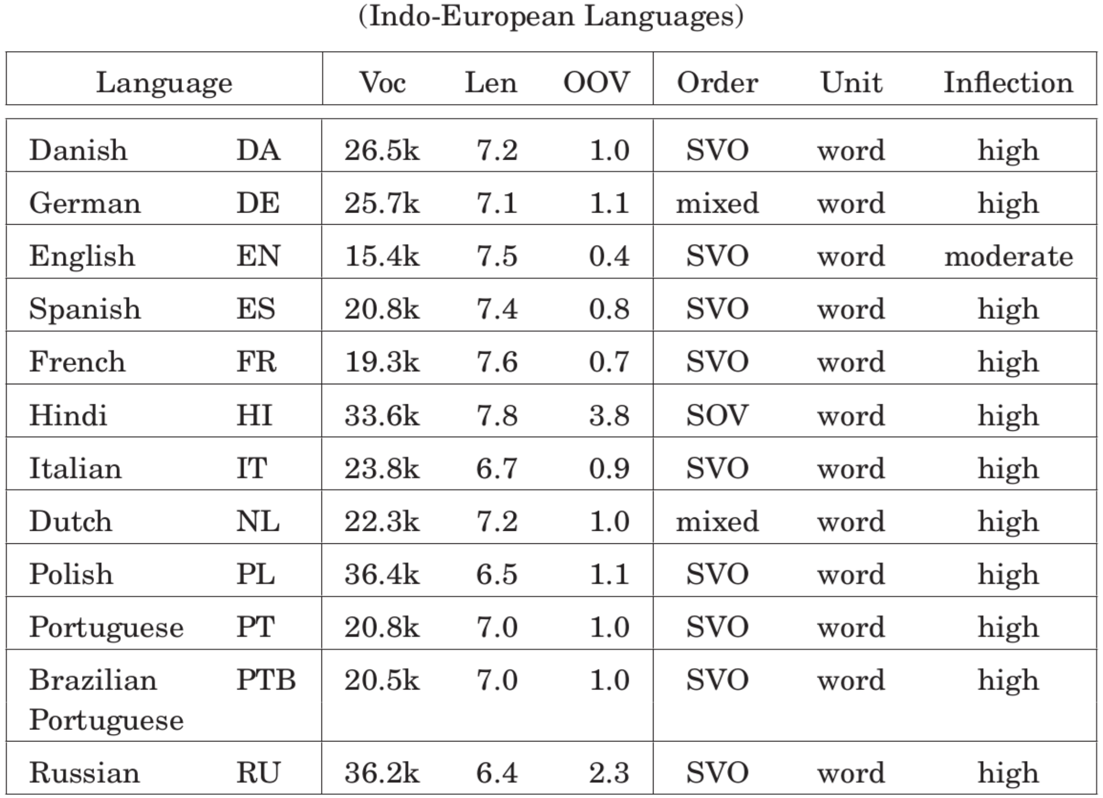
(1): Priority (9:1 ratio for Direct:Bridge table), (2) Priority by BLEU score



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**

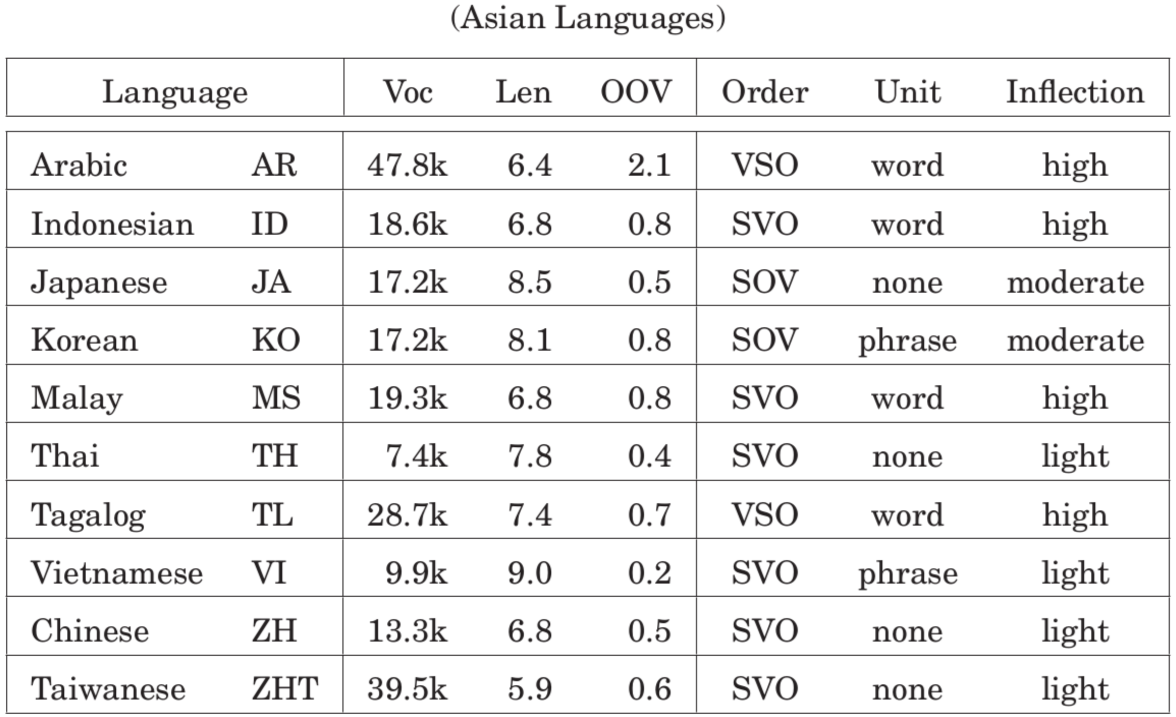
Japanese-Hindi translation using various pivots

Case Study VI *(Paul et al., 2013)*



(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

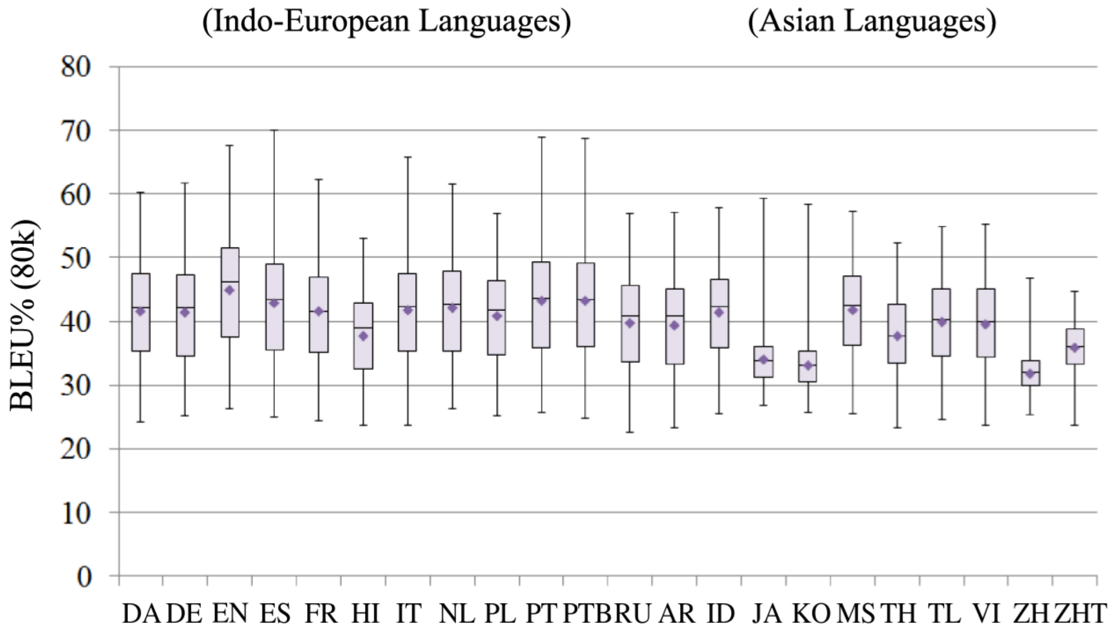


(Asian Languages)

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

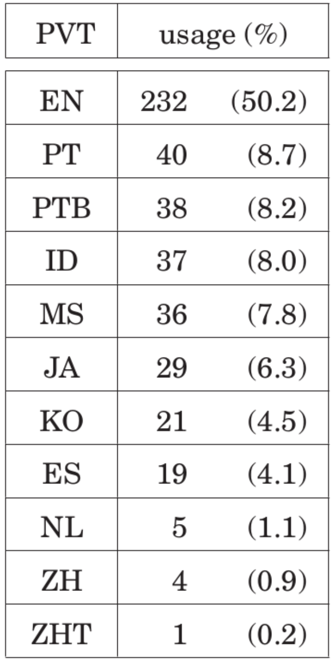
Study Involving 22 diverse Europoean and Asian languages

Case Study VI *(Paul et al., 2013)*



Pivot language dependency.

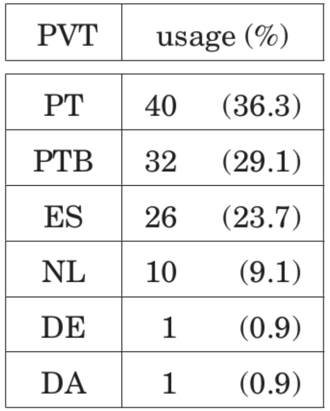
(All Languages)



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)

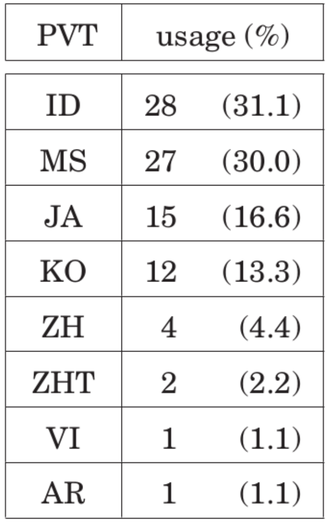
Non-English pivots

(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)

- **There is no single "best" pivot language**
- English good for 50.2% of language pairs

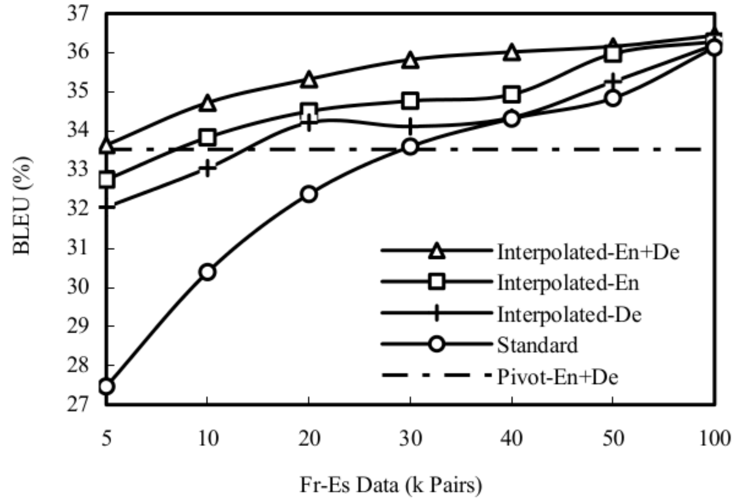
Closely related languages are generally good pivots

86% cases pivot language independent of data size

What if we use Multiple Pivots ?

Fr-Es translation using 2 pivots

Source: Wu & Wang (2007)



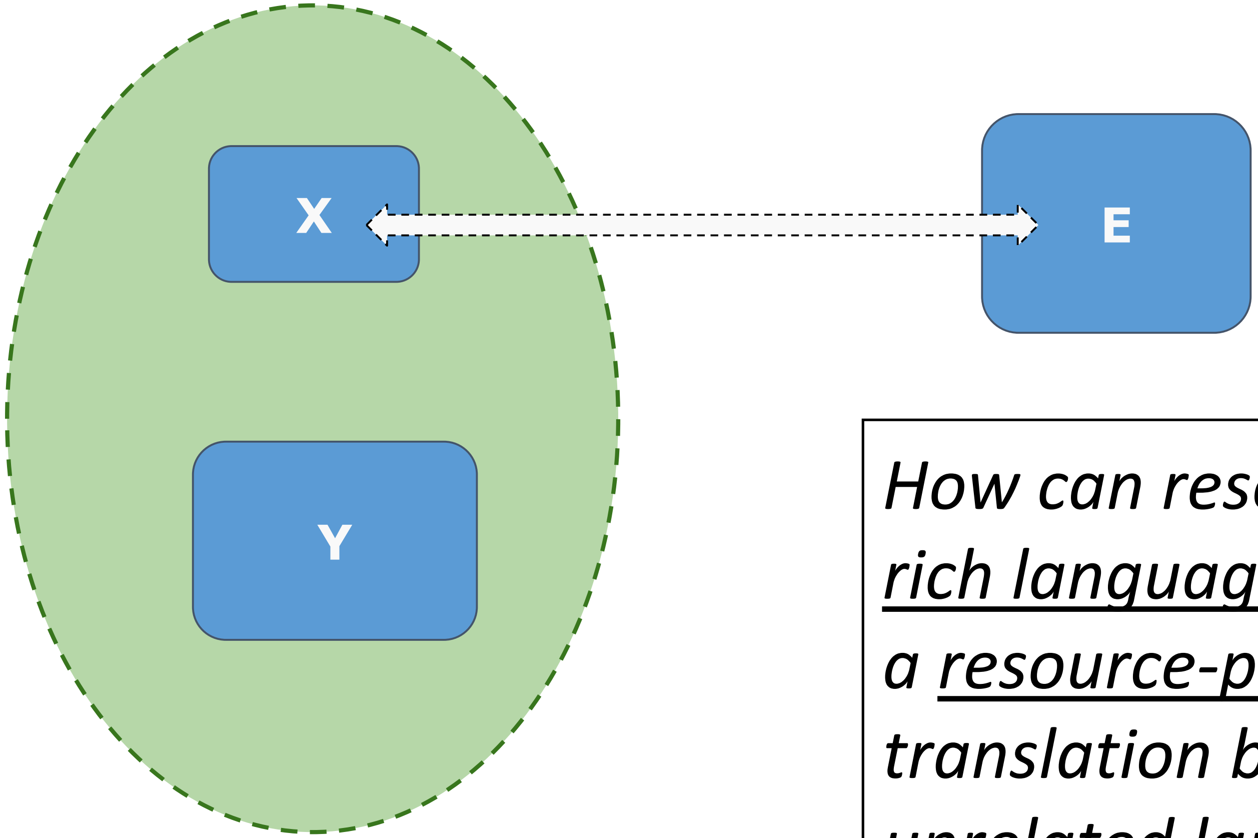
Hi ←→ Ja translation using 7 pivots

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

The more the better, especially when the training corpora are small

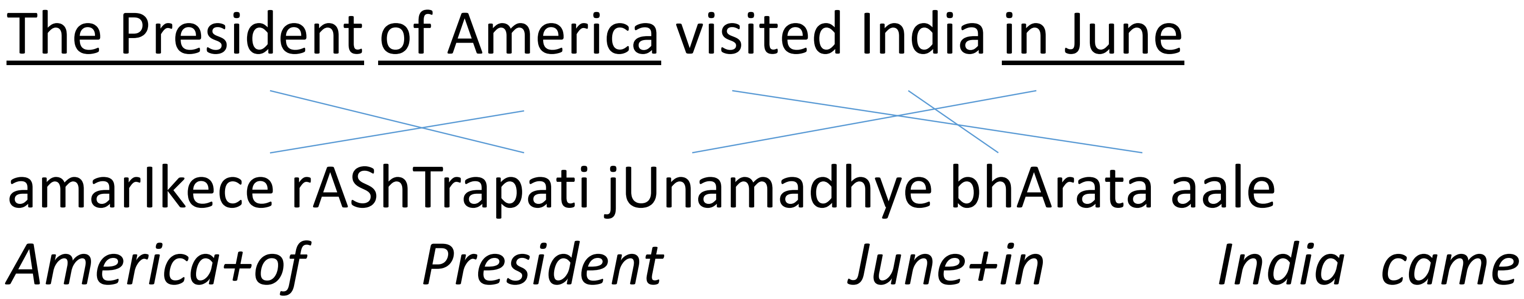


How can resources for a resource-rich language Y, which is related to a resource-poor language X, help translation between X, and an unrelated language E?

Can Y also help reduce structural divergence between X and E?

Consider English → Marathi translation

Word order divergence: English is SVO Marathi is SOV



The President of America visited India in June

amarIkece rAShTrapati jUnamadhye bhArata aale

America+of President June+in India came

Hindi and Marathi are Indo-Aryan languages with the same word order

Dravidian languages also have the same word order

Can reordering solutions for English → Hindi translation be reused for:

English → Marathi translation?

English → Telugu translation ?

Source Reordering

- Standard PBSMT cannot handle long-distance reordering
- *Source Reordering*: Change the word order of source side of the training corpus to match the target language word order prior to SMT training

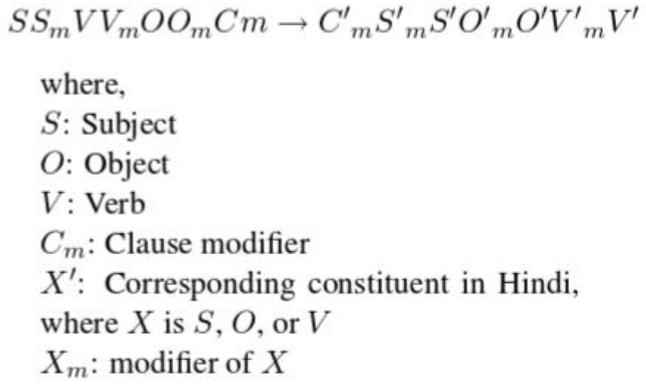
English	The President of America visited India in June
Reordered	America of The President June in India visited
Marathi	amarIkece rAShTrapati jUnamadhye bhAratat aale *America+of President June+in India+to came*

- 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→ Indian language translation

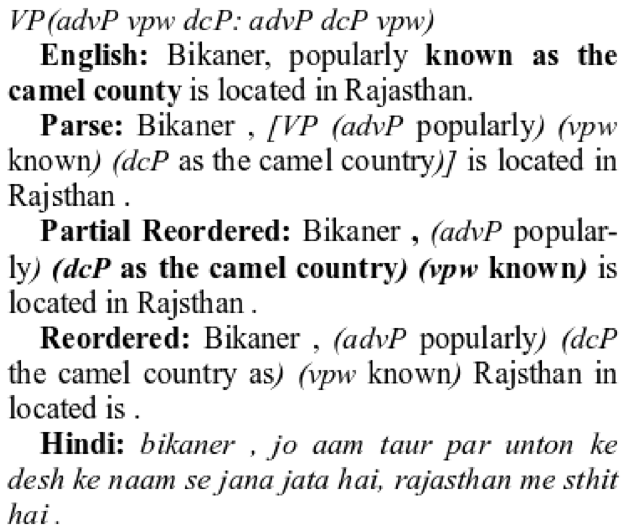


$$SS_mVV_mOO_mCm \rightarrow C'_mS'_mS'O'_mO'V'_mV'$$

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

(Kunchukuttan, et al. 2014)

	Indo-Aryan							Dravidian		
System	**hin**	**urd**	**pan**	**ben**	**guj**	**mar**	**kok**	**tam**	**tel**	**mal**
PBSMT	26.53	18.07	22.86	14.85	17.36	10.17	13.01	4.17	6.43	4.85
+generic reordering (S_1)	29.63	20.42	26.06	16.85	20.11	11.46	15.01	4.97	7.83	5.53
+Hindi tuned reordering (S_2)	30.86	21.54	27.52	18.20	21.33	12.68	15.73	5.09	8.29	5.68

- Source reordering improves BLEU scores for 15% and 21% for source reordering system systems S_1 and S_2 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

Tutorial Outline

- Introduction & Motivation
- Language Relatedness
- Translation within related languages
- Translation from related languages to another language
- **Summary**

Summary

We discussed the following questions …

- What is language relatedness & when is it useful for MT?
- Can translation between related languages be made more accurate?
- Can multiple languages help each other in translation?
- Can we reduce resource requirements?
- What concepts & tools are required for solving the above questions?

What does it mean to say languages are related?

- Genetic relation → Language Families
- Contact relation → Linguistic Area
- Linguistic typology → Linguistic Universals

Key characteristics of related languages

Lexical Similarity

Morphological correspondence

Monotonic word-order

Leverage these similarities to

→ *improve translation quality*
→ *reduce resource requirements*

Orthographic and Phonetic Similarity to measure word similarity

Properties & similarities of the scripts involved useful for measuring orthographic similarity

Identification of loan words, cognates, false friends and named entity pairs

Translation between related languages

- Adapting word-level SMT to improve word alignments, lexical coverage, OOV handling

- Use sub-word level units of representation

- Implicit use of morphological correspondence and monotonic word order

- Assistance from multiple languages via use of pivot languages

Food for thought

- Translation between related languages is not just transliteration *(Tsvetkov etal., 2015; Tsvetkov & Dyer, 2015)*
- Relation between lexical similarity and translation accuracy
- Evaluation Metrics for sub-word level transformations

Translation between related languages & another language

- Assisting language to improve vocabulary coverage & translation confidence
- Pivot based SMT to use corpus from a resource rich related language
- Source/Target rewriting: useful for related languages with little corpora
- Divergence between languages has to bridged
- *Linguistic resources can be re-used among related languages*

Can we reduce resource requirements?

- Lesser parallel corpora required for learning sub-word transformations
- Shared representation can be a powerful mechanism
- Resources can be re-used/ported between related languages

Key Tools and Concepts

- Language Typology
- Phonetic Properties
- Phonetic & Orthographic similarity
- Cognate Identification
- Transliteration

- Confusion networks & Word lattices representations
- Pivot-based MT
- Combining SMT models/outputs

Related Work that might be of interest

- Study of Linguistic Typology
- Historical/Comparative Linguistics
- Mining bilingual dictionaries, named entities & parallel corpora
- Word alignment using bridge languages
- Rule-based and Example-based MT in the light of linguistic similarities
- Multilingual Neural Machine Translation
- Character-level Neural Machine Translation

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

Classification of Reading Material

Language Relatedness:	1,7,15,16,49,53,56
Lexical Similarity:	9,18,20,22,23,24,29,31,32,46,63
Adapting word-level SMT	8,12,13,14,24,26,28,34,41,47,55,59,60
Character-level SMT	36,37,38,39,57,58
Pivot-based SMT	5,10,11,25,30,33,35,37,43,44,61,62,64,65,66

List of papers at the end

Thank You!

Questions?

References

1. 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
2. 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
3. Emily Bender. *On achieving and evaluating language-independence in NLP*. Linguistic Issues in Language Technology. 2011.
4. Shane Bergsma, Grzegorz Kondrak. *Alignment-based discriminative string similarity*. Annual meeting-Association for Computational Linguistics. 2007.
5. N. Bertoldi, M. Barbaiani, M. Federico, R. Cattoni. *Phrase-based statistical machine translation with pivot languages*. IWSLT. 2008.
6. 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.
7. Peter Daniels and William Bright. *The world's writing systems*. Oxford University Press, 1996.
8. 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.
9. Michael Covington. *An algorithm to align words for historical comparison*. Computational linguistics. 1996.
10. 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.
11. 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.

References

12. Nadir Durrani, Hassan Sajjad, Hieu Hoang and Philipp Koehn. *Integrating an unsupervised transliteration model into statistical machine translation*. EACL. 2014.

13. 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.

14. 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.

15. 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/

16. Murray Emeneau. *India as a linguistic area*. Language. 1956.

17. 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.

18. 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.

19. 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.

20. 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.

21. Philipp Koehn. *Statistical machine translation*. Cambridge University Press. 2009.

References

22. Greg Kondrak. *Cognates and word alignment in bitexts*. MT Summit. 2005.

23. 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.

24. 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.

25. S. Kumar, Och, F. J., Macherey, W. *Improving word alignment with bridge language*s. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007.

26. 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.

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

28. Anoop Kunchukuttan, Ratish Pudupully, Rajen Chatterjee, Abhijit Mishra, Pushpak Bhattacharyya. 2014. *The IIT Bombay SMT System for ICON 2014 Tools Contest* . NLP Tools Contest at ICON 2014.

29. 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.

30. Evgeny Matusov, Nicola Ueffing, and Hermann Ney. *Computing Consensus Translation for Multiple Machine Translation Systems Using Enhanced Hypothesis Alignment.* EACL. 2006.

References

31. Dan Melamed. *Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons*. Third Workshop on Very Large Corpora. 1995.
32. Dan Melamed. *Models of translational equivalence among words*. Computational Linguistics. 2000.
33. Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura. *Improving Pivot Translation by Remembering the Pivo*t. Association for Computational Linguistics. 2015.
34. 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.
35. Rohit More, Anoop Kunchukuttan, Raj Dabre, Pushpak Bhattacharyya. *Augmenting Pivot based SMT with word segmentation*. International Conference on Natural Language Processing. 2015.
36. 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.
37. 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.
38. 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.
39. Graham Neubig, Taro Watanabe, Shinsuke Mori, Tatsuya Kawahara. *Machine Translation without Words through Substring Alignment.* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012.

References

40. Franz Och and Hermann Ney. *Statistical multi-source translation*. In Proceedings of MT Summit VIII. Machine Translation in the Information Age , MT Summit. 2001.

41. Franz Och, and Hermann Ney. *A systematic comparison of various statistical alignment models*." Computational linguistics. 2003.

42. Raj Nath Patel, Rohit Gupta, and Prakash B. Pimpale. *Reordering rules for English-Hindi SMT*. HYTRA. 2013.

43. 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.

44. 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.

45. 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.

46. E. Ristad, P. Yianilos. *Learning string-edit distance.* IEEE Trans. Pattern Anal. Mach. Intell., 20(5):522–532, 1998.

47. 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.

48. 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.

49. 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.

References

50. 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.

51. 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.

52. David Steele, Lucia Specia. *WA-Continuum: Visualising Word Alignments across Multiple Parallel Sentences Simultaneously*. ACL-IJCNLP. 2015.

53. Karumuri Subbarao. *South Asian languages : a syntactic typology*. Cambridge University Press. 2012.

54. 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.

55. 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.

56. Sarah Thomason. *Linguistic Areas and Language History*. Studies in Slavic and General Linguistics. 2000.

57. Jorge Tiedemann. *Character-based PSMT for closely related languages*. In Proceedings of the 13th Annual Conference of the European Association for Machine Translation. 2009.

58. 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.

59. Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. *Constraint-Based Models of Lexical Borrowing***.** In *Proc. NAACL'15.*

References

60. Yulia Tsvetkov and Chris Dyer. Lexicon Stratification for Translating Out-of-Vocabulary Words. In Proc. ACL'15.

61. Masao Utiyama, Hitoshi Isahara. A *comparison of pivot methods for phrase-based statistical machine translation.* In HLT-NAACL, pages 484–491, 2007.

62. D. Vilar, Peter, J.-T., & Ney, H.. *Can we translate letters?.* In Proceedings of the Second Workshop on Statistical Machine Translation. 2007.

63. Robert Wagner, Michael J. Fischer. *The string-to-string correction problem*. Journal of the ACM. 1974.

64. Haifeng Wang, Hua Wu, and Zhanyi Liu. *Word alignment for languages with scarce resources using bilingual corpora of other language pairs.* COLING-ACL. 2006.

65. Hua Wu, Haifeng Wang. *Pivot language approach for phrase-based statistical machine translation.* Machine Translation. 2007.

66. Robert Östling. *Bayesian word alignment for massively parallel texts*. 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014.