Multilingual Learning

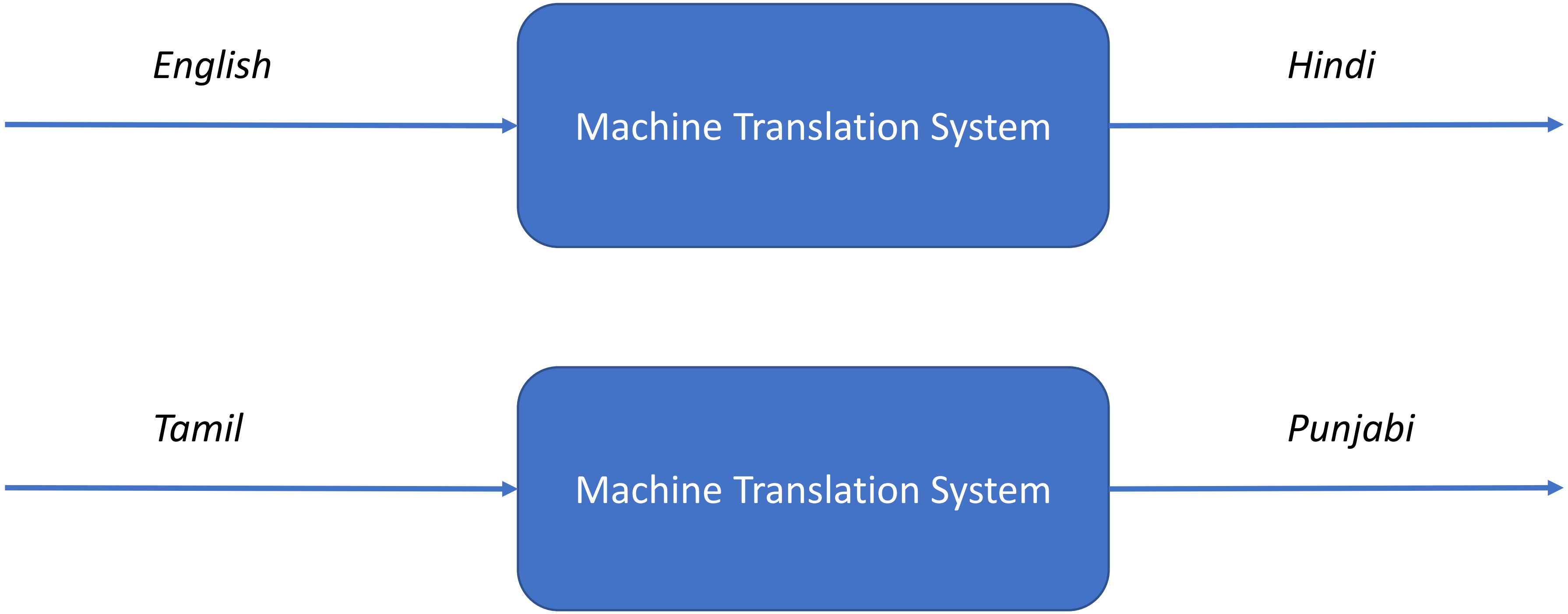
Anoop Kunchukuttan

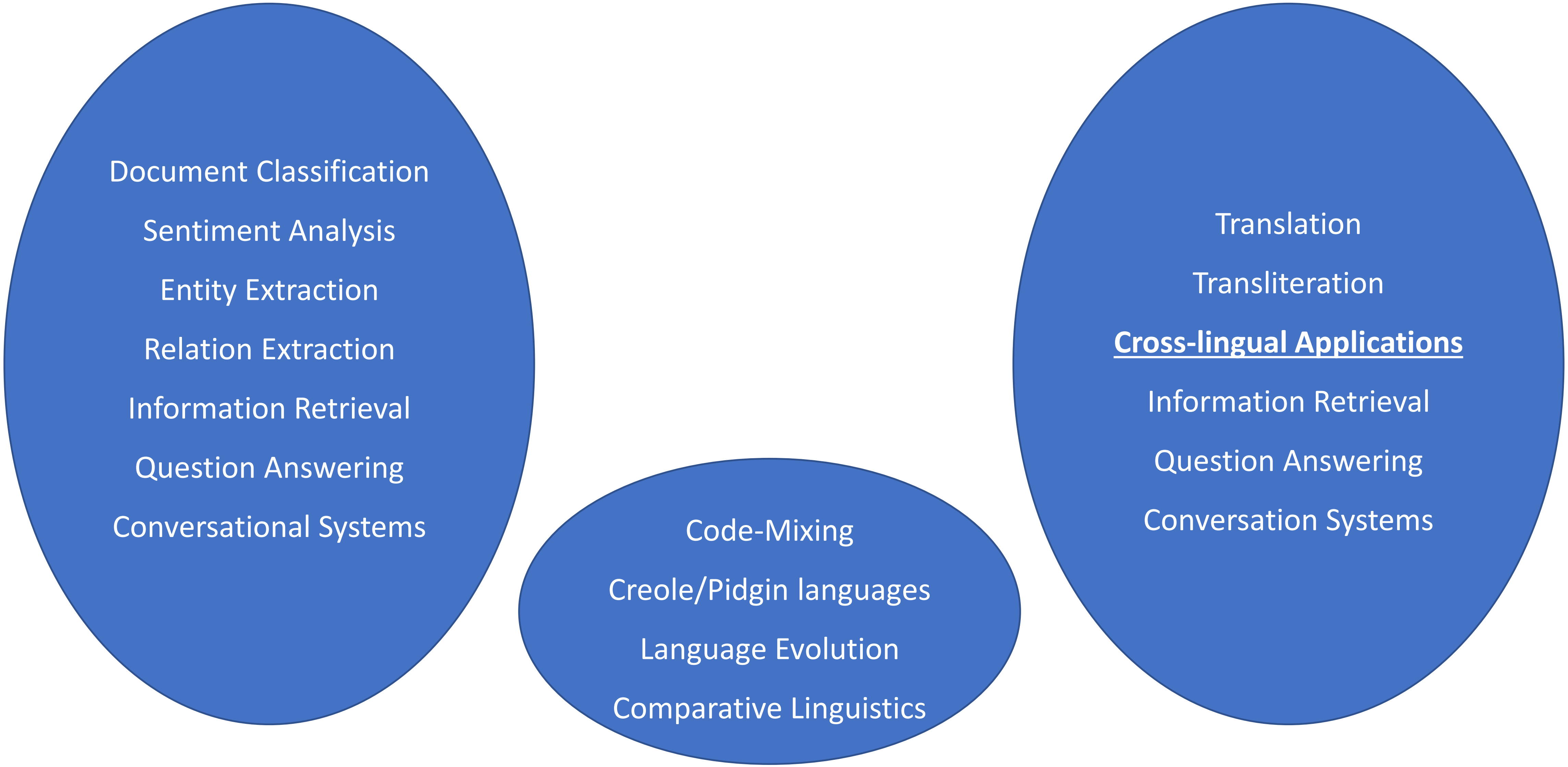
Microsoft AI and Research

Center for Indian Language Technology
Indian Institute of Technology Bombay

3rd Summer School on Machine Learning (Advances in Modern AI), 13th July 2018

Broad Goal: Build NLP Applications that can work on different languages





Monolingual Applications

Document Classification

Sentiment Analysis

Entity Extraction

Relation Extraction

Information Retrieval

Question Answering

Conversational Systems

Cross-lingual Applications

Translation

Transliteration

Cross-lingual Applications

Information Retrieval

Question Answering

Conversation Systems

Code-Mixing

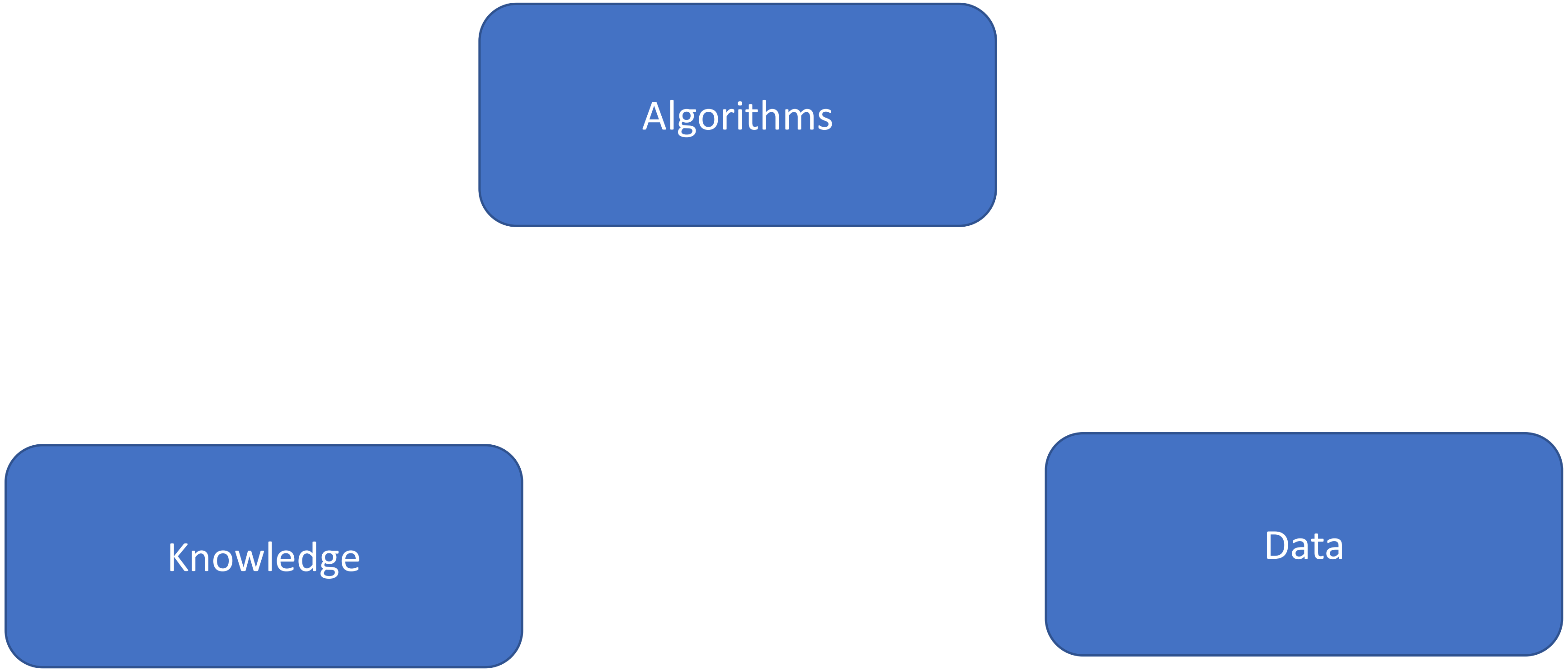
Creole/Pidgin languages

Language Evolution

Comparative Linguistics

Mixed Language Applications

Facets of an NLP Application



Algorithms

Knowledge

Data

Facets of an NLP Application

RULE-BASED SYSTEMS

Algorithms

Expert Systems
Theorem Provers
Parsers
Finite State Transducers

Largely language independent

Knowledge

Rules for morphological analyzers, Production rules, etc.

Lot of linguistic knowledge encoded

Data

Paradigm Tables, dictionaries, etc.

Lot of linguistic knowledge encoded

Some degree of language independence through good software engineering and knowledge of linguistic regularities

Facets of an NLP Application

STATISTICAL ML SYSTEMS (Pre-Deep Learning)

Algorithms

Largely language independent, could solve non-trivial problems efficiently

Supervised Classifiers
Sequence Learning Algorithms
Probabilistic Parsers
Weighted Finite State Transducers

Knowledge

Feature Engineering

Lot of linguistic knowledge encoded
Feature engineering is easier than maintain rules and knowledge-bases

Data

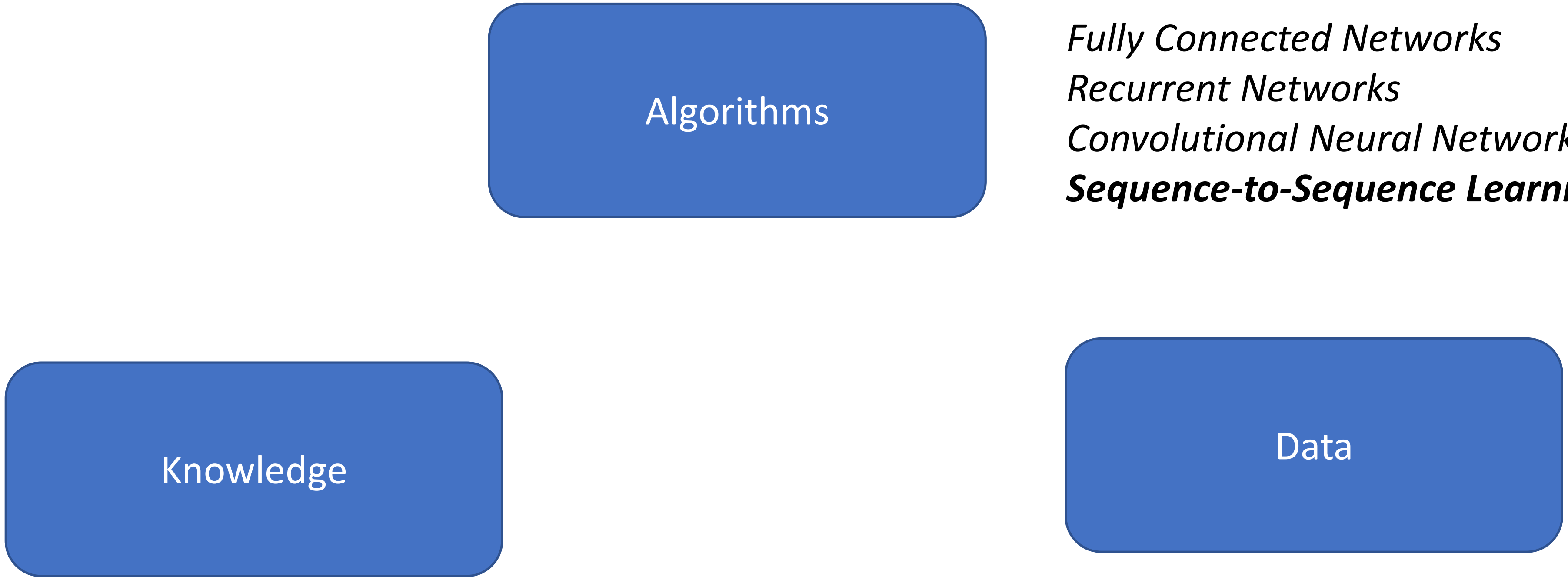
***Annotated Data**, Paradigm Tables, dictionaries, etc.*

Lot of linguistic knowledge encoded

General language-independent ML algorithms and easy feature learning

Facets of an NLP Application

DEEP LEARNING SYSTEMS



Largely language independent

Algorithms

Fully Connected Networks
Recurrent Networks
Convolutional Neural Networks
Sequence-to-Sequence Learning

Knowledge

Data

Representation Learning, *Architecture Engineering, AutoML*

Annotated Data*~~, Paradigm Tables, dictionaries, etc.~~*

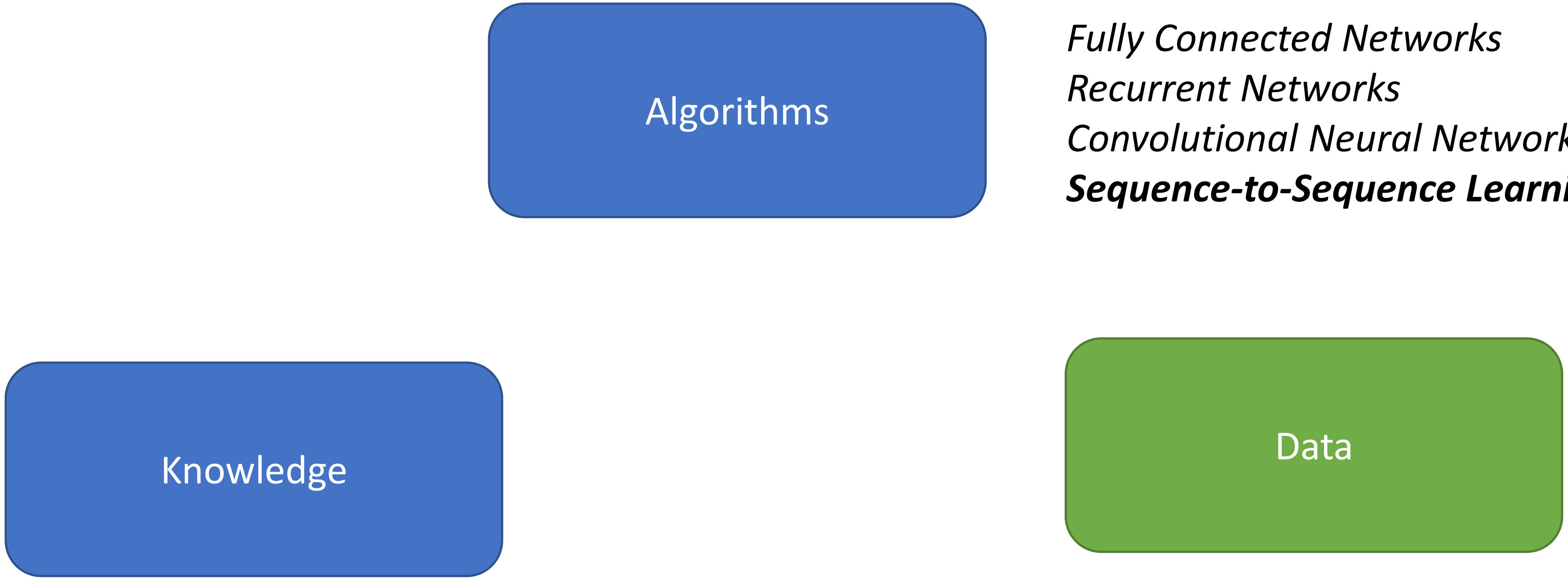
Very little knowledge; annotated data is still required

Feature engineering is unsupervised, largely language independent

Neural Networks provide a convenient language for expressing problems, representation learning automated feature engineering

Facets of an NLP Application

DEEP LEARNING SYSTEMS



Largely language independent

Fully Connected Networks
Recurrent Networks
Convolutional Neural Networks
Sequence-to-Sequence Learning

Representation Learning, *Architecture Engineering, AutoML*

Annotated Data*~~, Paradigm Tables, dictionaries, etc.~~*

Very little knowledge; annotated data is still required

Feature engineering is unsupervised, largely language independent

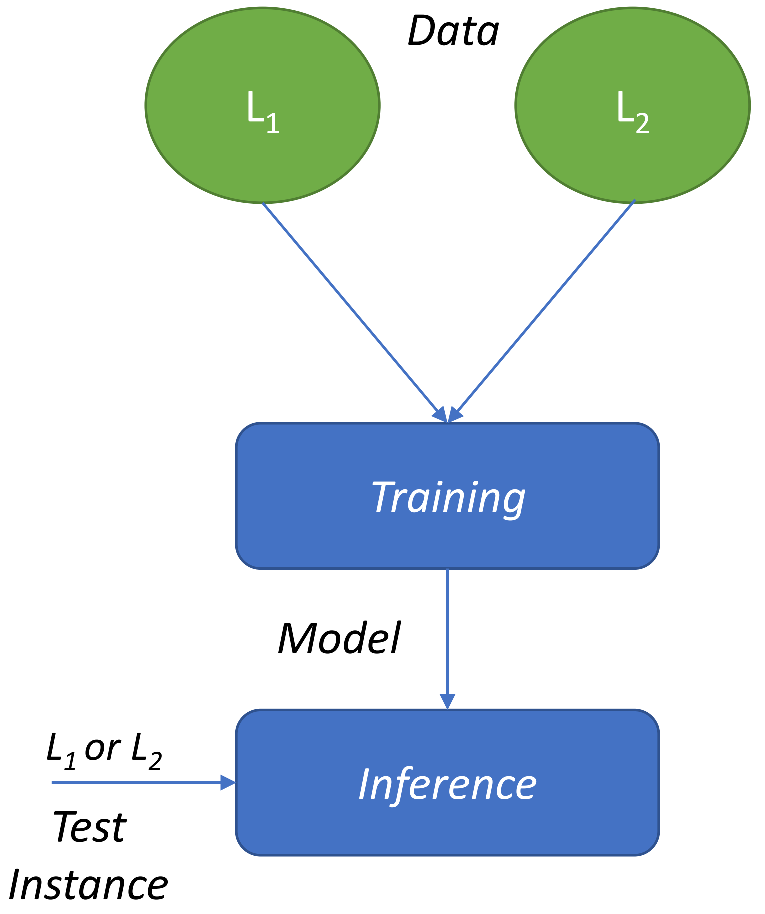
Neural Networks provide a convenient language for expressing problems, representation learning automated feature engineering

Focus of today’s session

How to leverage data for one language to build NLP applications for another language?

Multilingual Learning Scenarios

Joint Learning

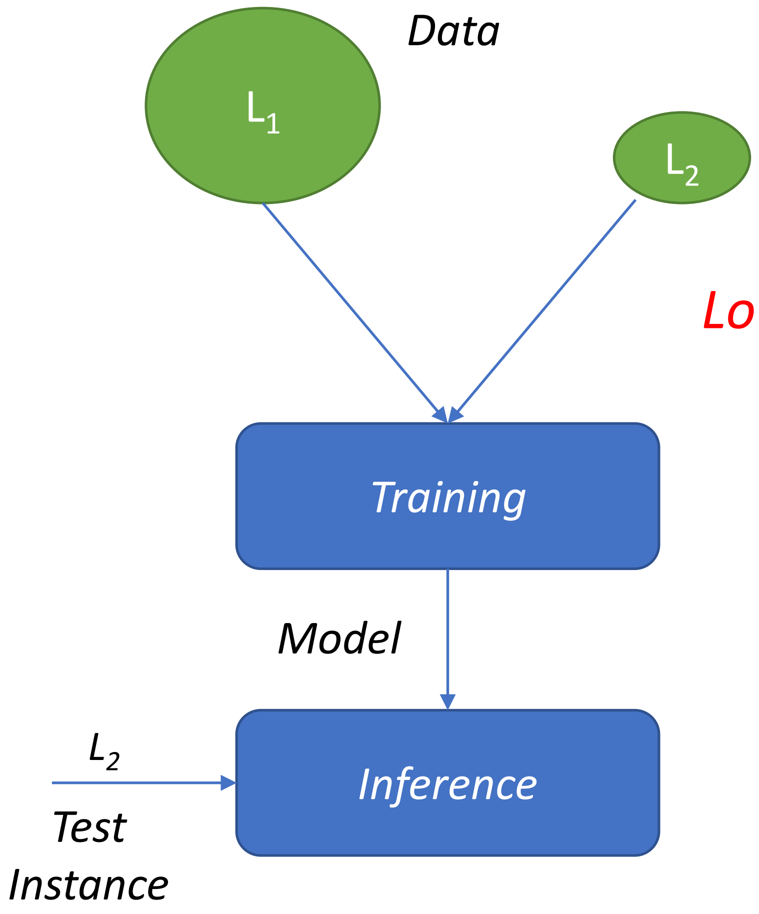


- *Analogy to Multi-task learning* ➔ *Task ≡ Language*
- *Related Tasks can share representations*
- *Representation Bias: Learn the task to generalize over multiple languages*
- *Eavsdropping*
- *Data Augmentation*

(Caruana., 1997)

Multilingual Learning Scenarios

Transfer Learning

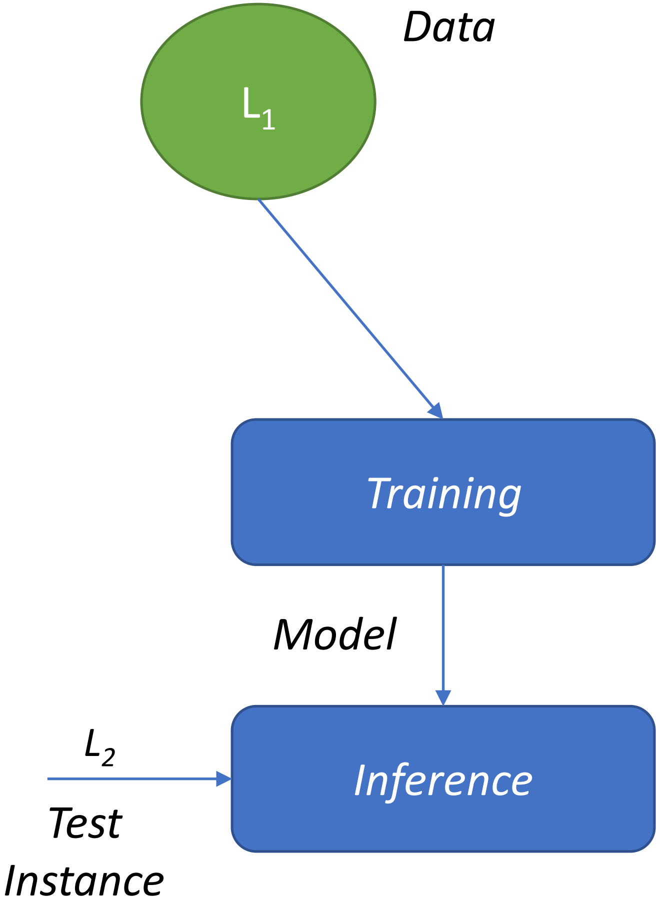


Low resource language can benefit from data for high resource language

(Caruana., 1997)

Multilingual Learning Scenarios

Zeroshot Learning

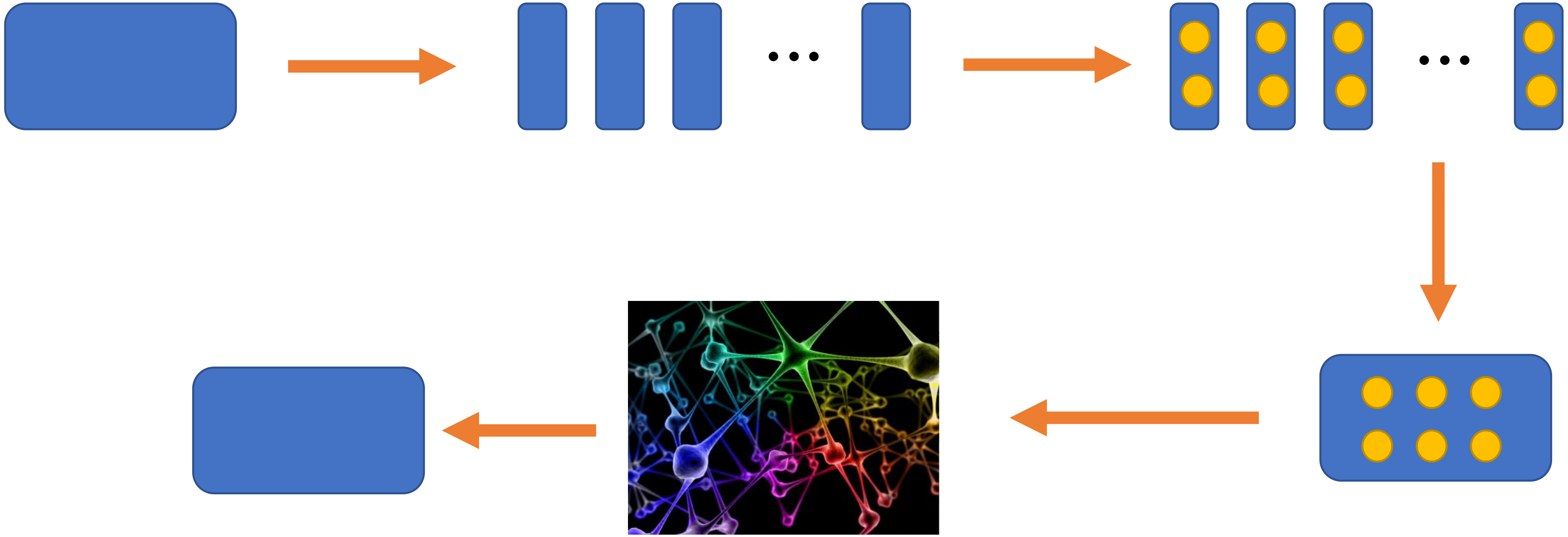


Can system be trained for one language so that they work out of the box for another language?

What does Deep Learning bring to the table?

- Neural Networks provide a powerful framework for Multilingual learning
    - *Caruana's seminal work on Multi-task learning in 1997 used Neural Networks*
- Word embeddings: Powerful feature representation mechanism to capture syntactic and semantic similarities
    - *Distributed representation*
    - *Unsupervised learning*
- Algebraic reasoning as opposed to Mathematical Logic
- Numerical optimization as opposed to combinatorial optimization

A Typical Multilingual NLP Pipeline



Text

Tokens

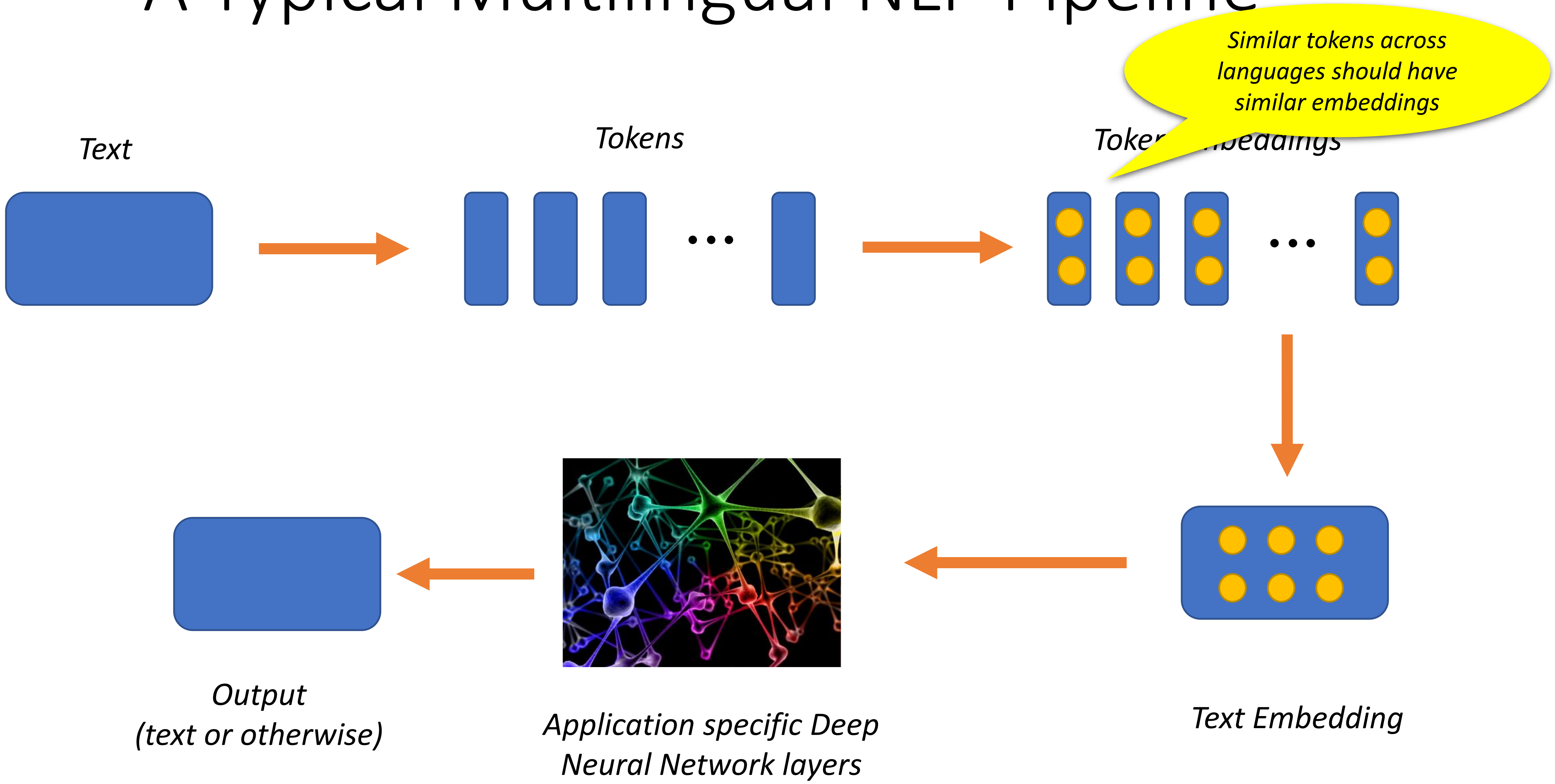
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

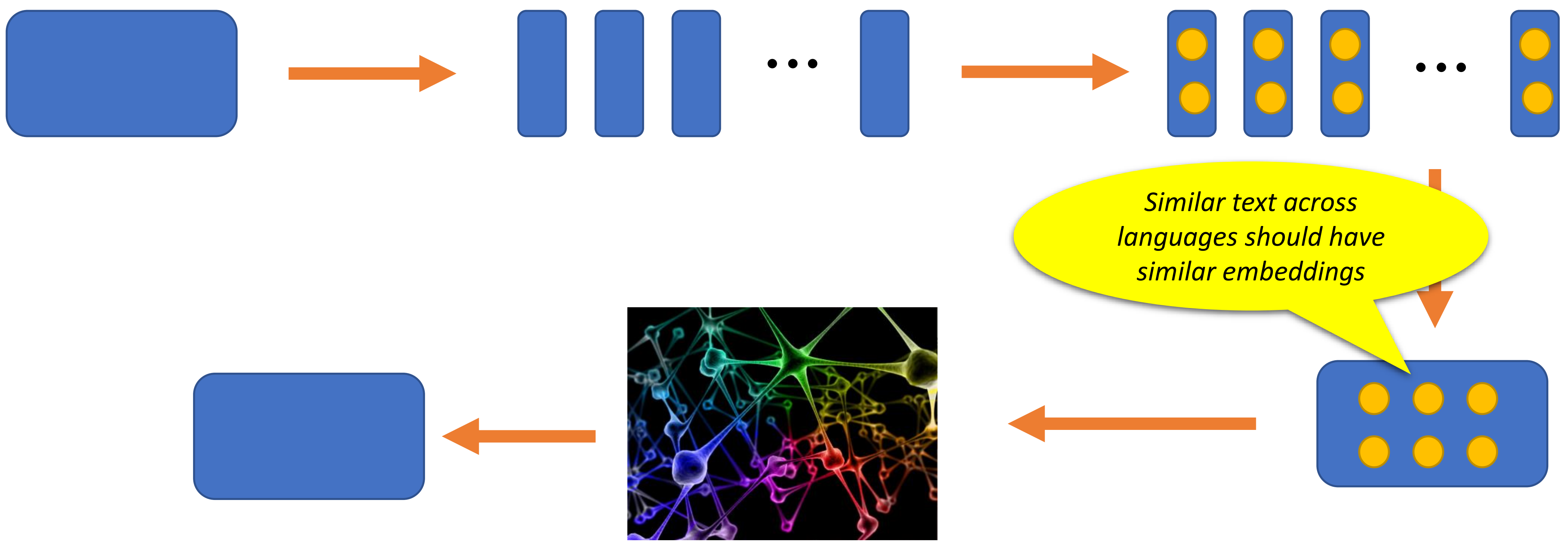
Similar tokens across languages should have similar embeddings

Text Embedding

Application specific Deep Neural Network layers

Output
(text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

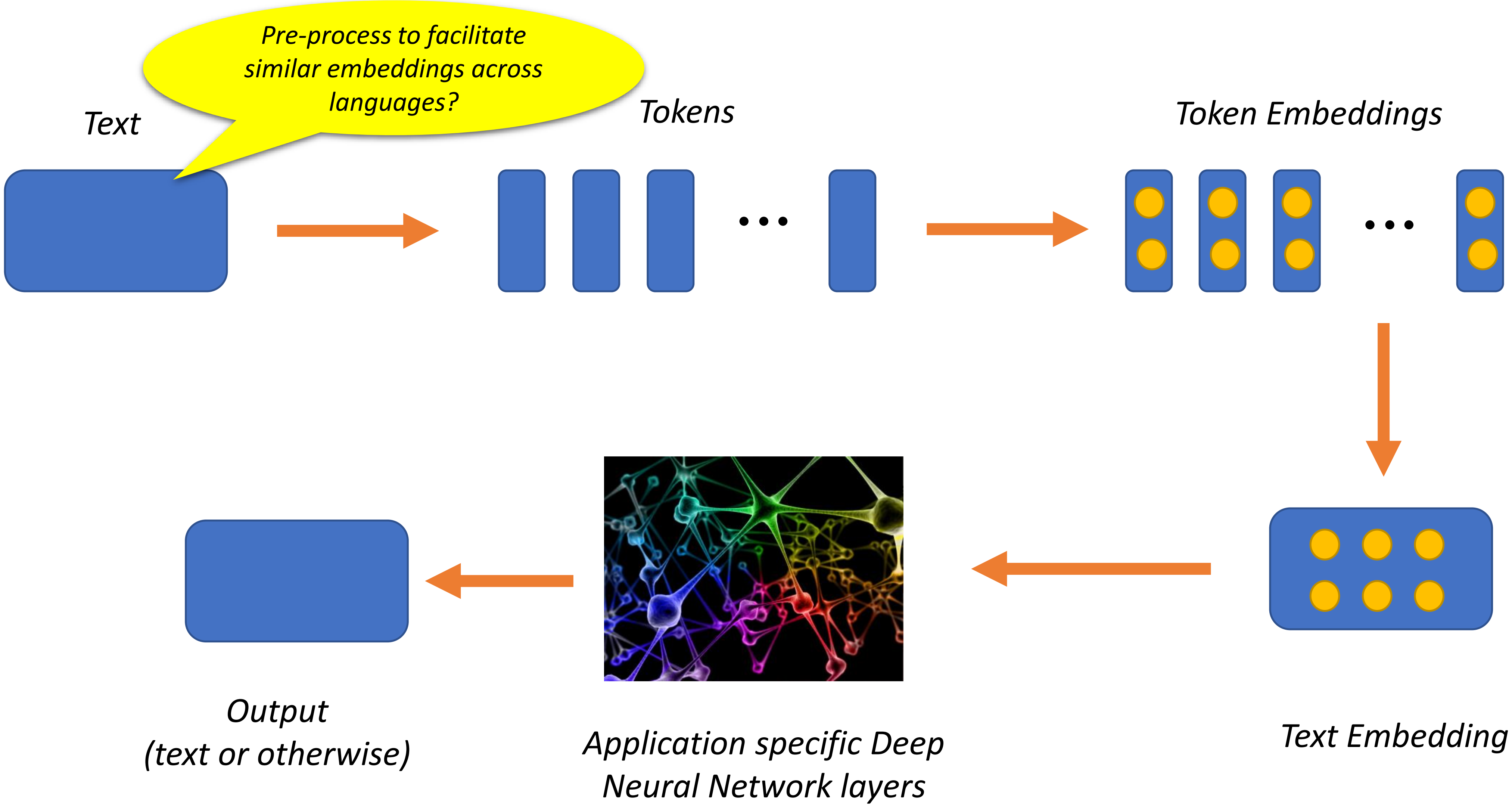
Similar text across languages should have similar embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Pre-process to facilitate similar embeddings across languages?

Text

Tokens

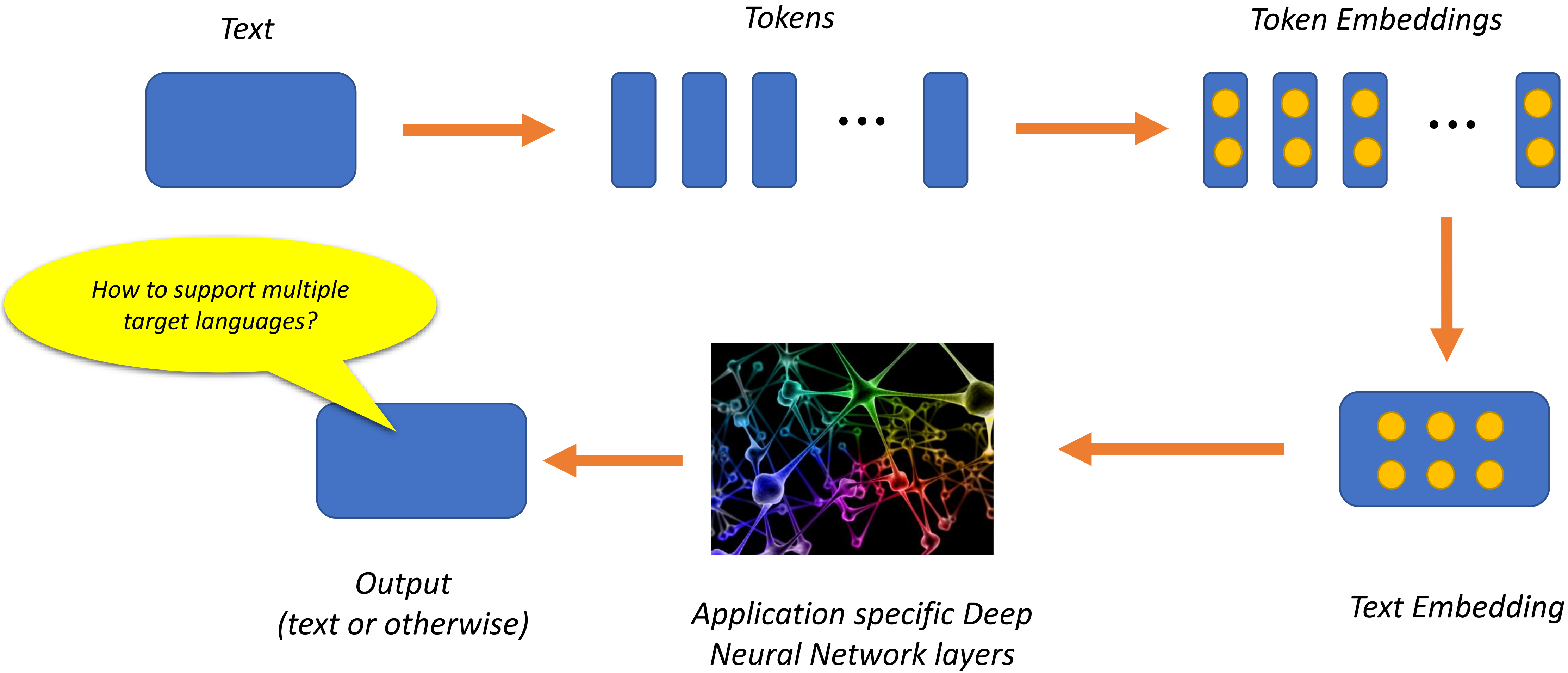
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

How to support multiple target languages?

Outline

- Learning Cross-lingual Embeddings

- Training a Multilingual NLP Application

- Related Languages and Multilingual Learning

- Summary and Research Directions

Cross-Lingual Embeddings

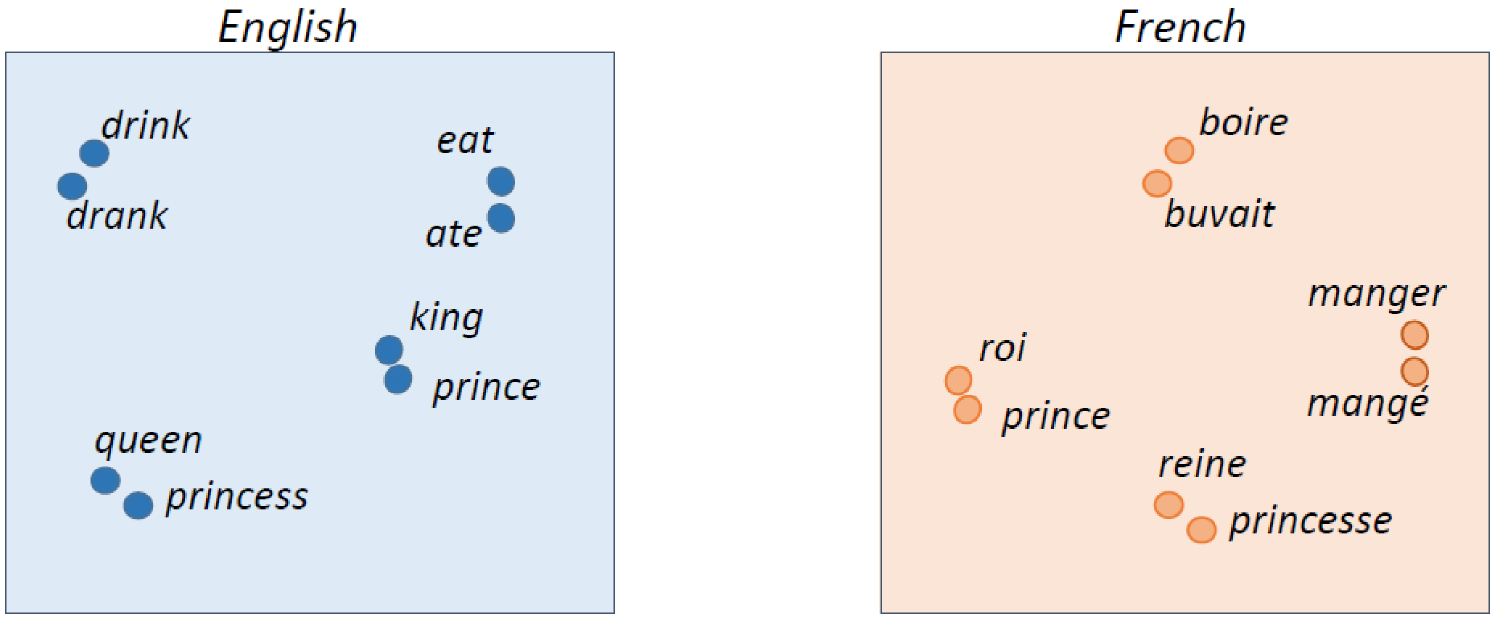
Offline Methods

Online Methods

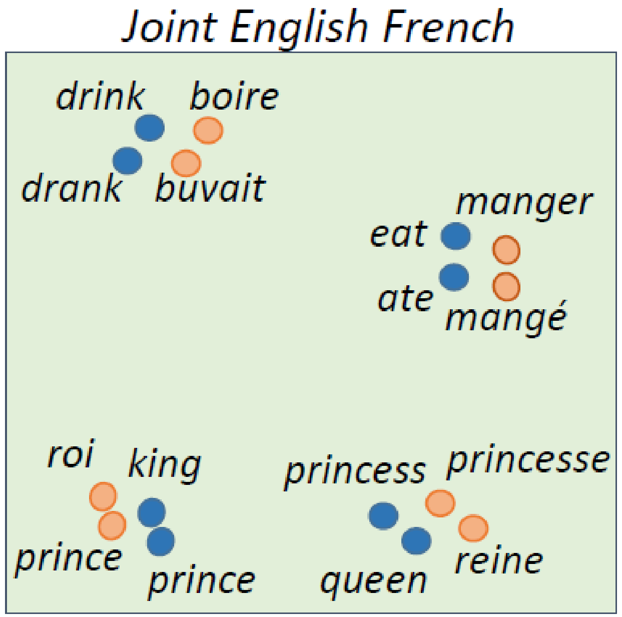
Some observations

Evaluation

Unsupervised Learning



<u>Monolingual Word Representations</u>
(capture syntactic and semantic similarities between words)



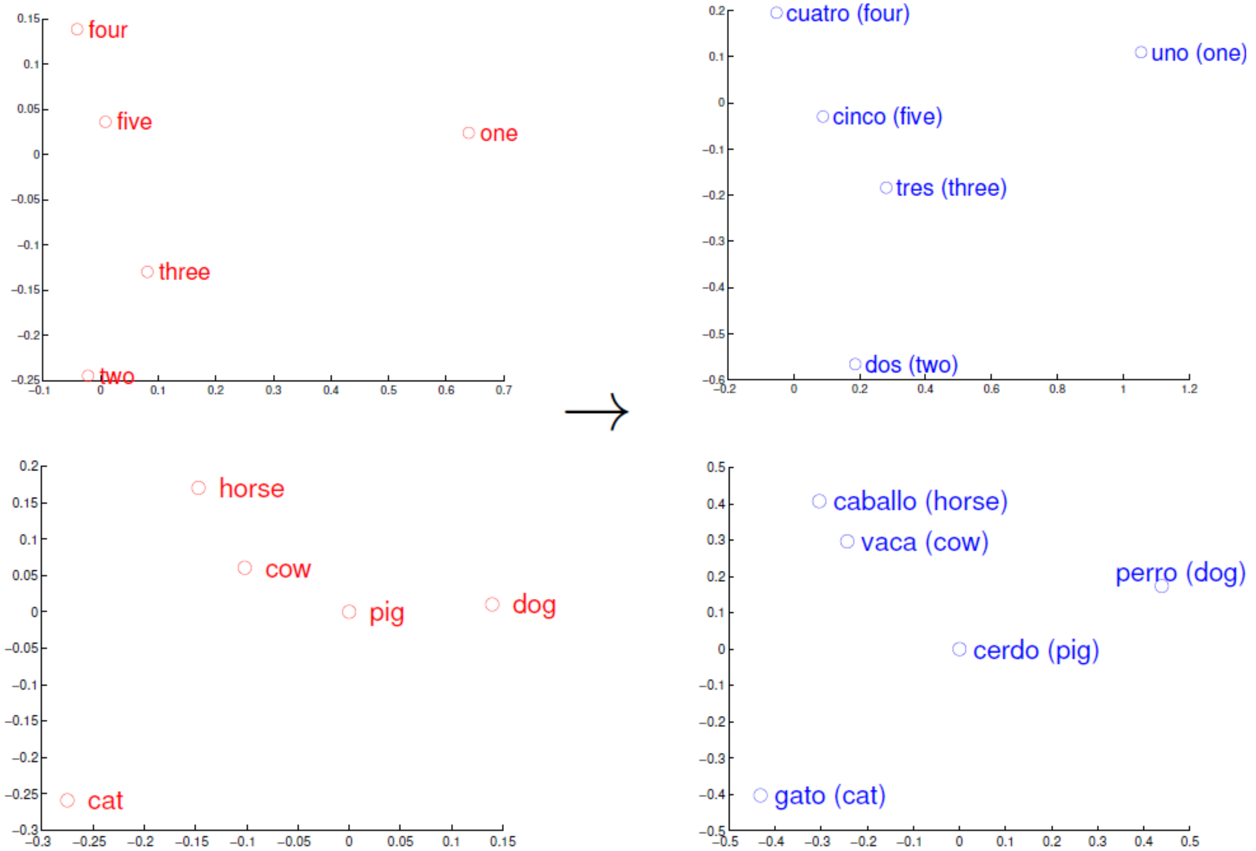
<u>Multilingual Word Representations</u>
(capture syntactic and semantic similarities between words both <u>within and across languages</u>)

(Source: Khapra and Chandar, 2016)

$$embed(y) = f(embed(x))$$

x, y are source and target words
$embed(w)$: embedding for word w

Is it possible to learn mapping functions?



- Languages share concepts ground in the real world
- Some evidence of universal semantic structure (*Youn et al., 2016)*
- Isomorphism between embedding spaces (*Mikolov et al., 2013)*
- Isomorphism can be captured via a linear transformation

(Source: Mikolov et al., 2013)

Offline Methods

Learn monolingual and cross-lingual embeddings separately

General require weaker parallel signals

e.g., bilingual dictionaries

Online Methods

Learn monolingual and cross-lingual embeddings jointly

Generally require stronger parallel signals

e.g., parallel corpus

Cross-Lingual Embeddings

Offline Methods

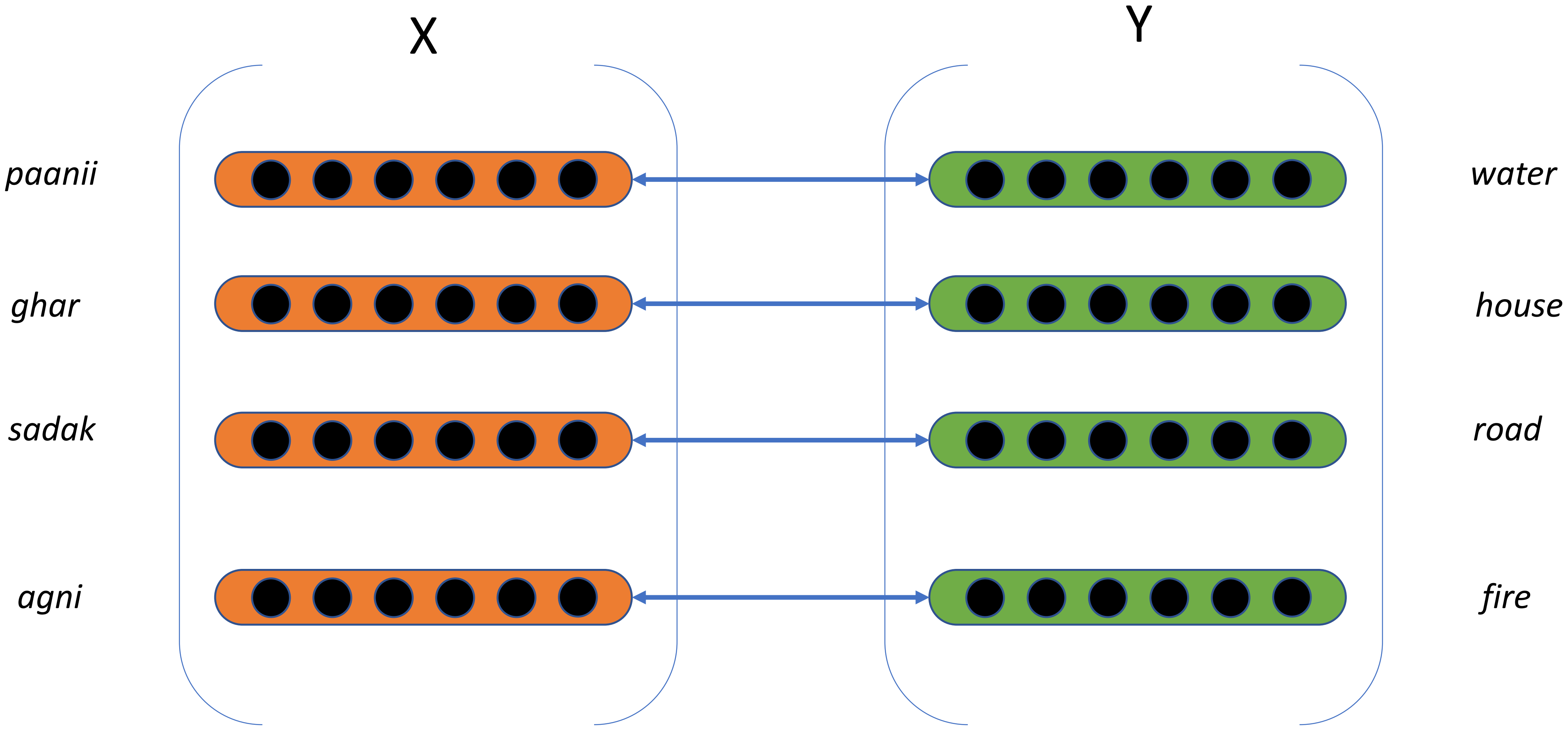
Online Methods

Some observations

Evaluation

Unsupervised Learning

Supervised Learning



$$XW = Y$$

Least Squares Solution

(Mikolov et al., 2013)

$$W^* = \underset{W \in \mathbb{R}^d}{\operatorname{argmin}} \|XW - Y\|_2^2$$

We can have a closed form solution:

$$X^+ = (X^T X)^{-1} X^T$$

$$W^* = X^+ Y$$

Solutions can be regularized using L_1*or* L_2 *norms to prevent overfitting*

Orthogonality Constraint on W

$$W^T W = I$$

- Preserves similarity in the target space *(Artetxe et al., 2016)*

$$(Wx)^T(Wy) = x^T W^T W y = x^T y$$

- Mapping Function is reversible *(Smith et al., 2017)*

$$W^T W x = x$$

- If source embeddings are unit vectors, orthogonality ensures target is also a unit vector *(Xing et al., 2015)*

$$y^T y = (Wx)^T(Wx) = x^T W^T W x = x^T x = 1$$

- Why length normalize? ➔ dot product equivalent to cosine similarity

Orthogonal Procrustes Problem

(Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017)

$$W^* = \underset{W \in O^d}{\operatorname{argmin}} \|XW - Y\|_2^2$$

We can have a closed form solution to this problem too (Schönemann, 1966)

$$Y^T X = U \Sigma V^T$$

$$W^* = V U^T$$

If embeddings are length-normalized, the above objective is equivalent to maximizing cosine similarity

$$W^* = \underset{W \in O^d}{\operatorname{argm}ax} \sum_i \cos(X_{i*} W, Y_{i*})$$

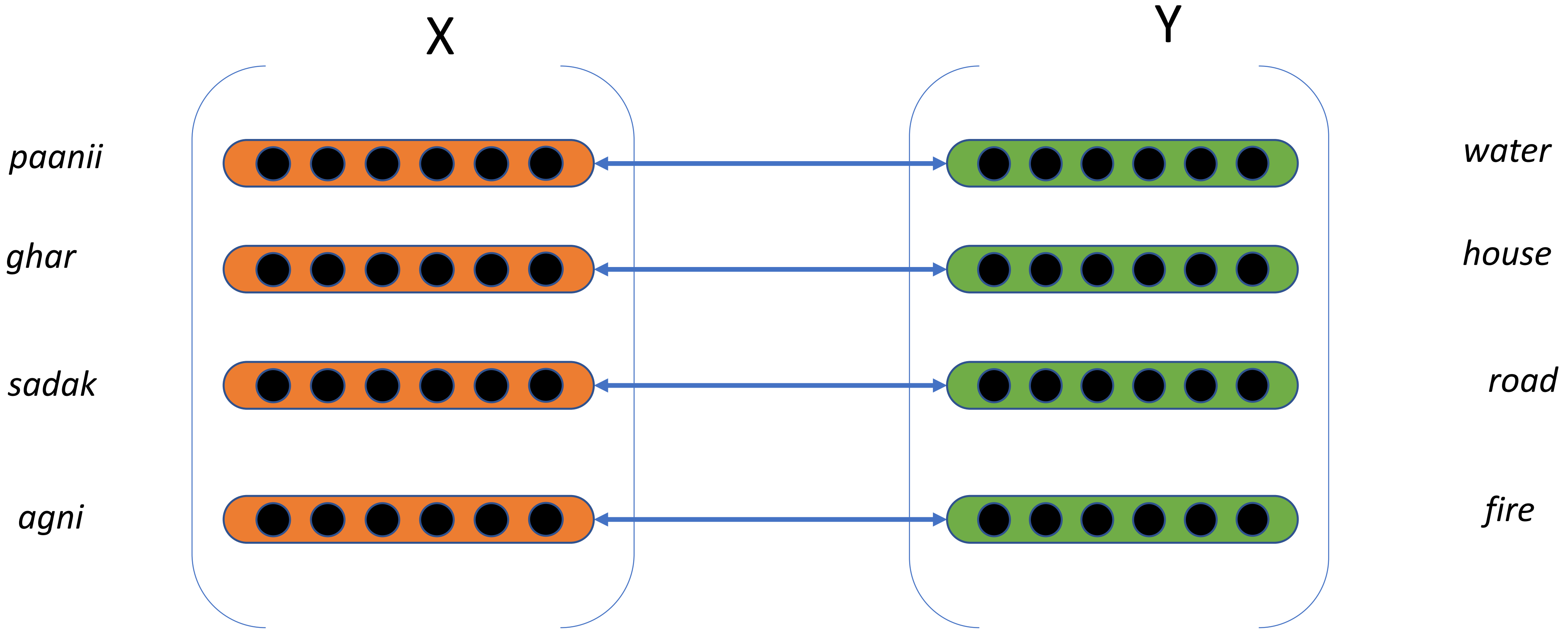
Canonical Correlation Analysis (CCA)

(Faruqui and Dyer, 2014; Ammar et al. 2015)

Regression methods → maximize similarity between target & mapped source embeddings

An alternative way to compare:

Is there a latent space where the dimensions of the embeddings are correlated?



X

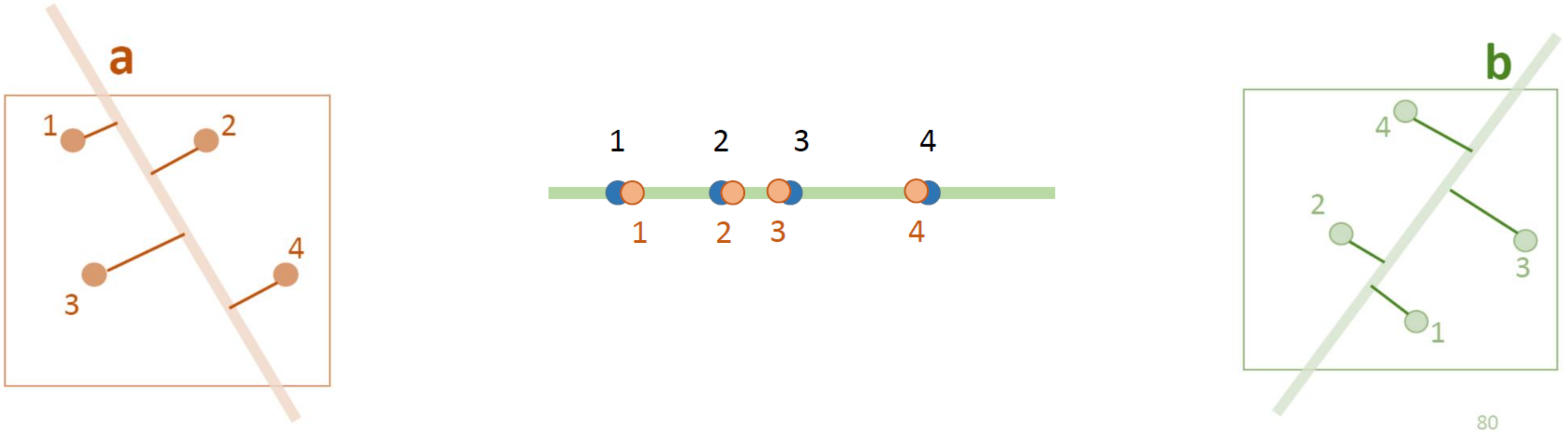
Y

paanii — *water*

ghar — *house*

sadak — *road*

agni — *fire*



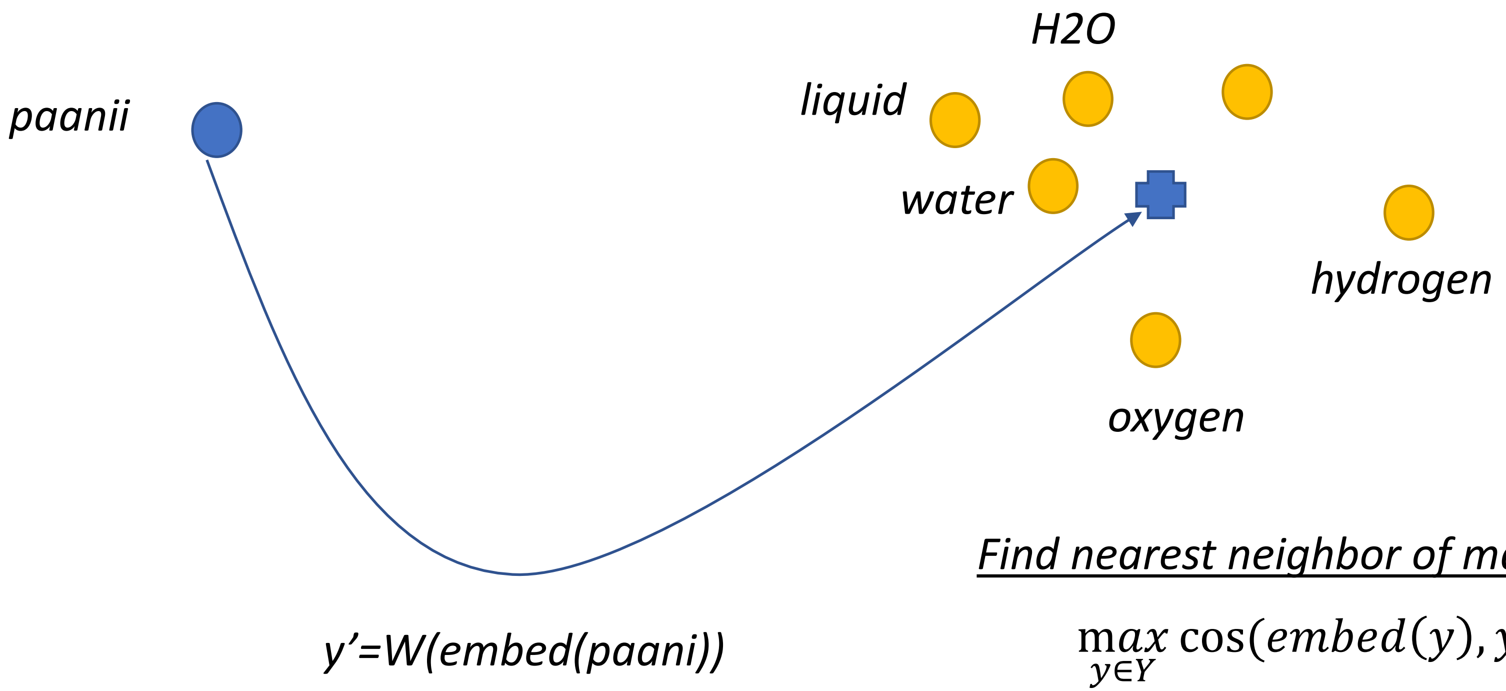
$$maximize\ trace((XA)^T(YB))$$

This term capture the correlation between the dimensions in the latent space defined by A and B

Bilingual Lexicon Induction

Given a mapping function and source/target words and embeddings:

Can we extract a bilingual dictionary?



Find nearest neighbor of mapped embedding

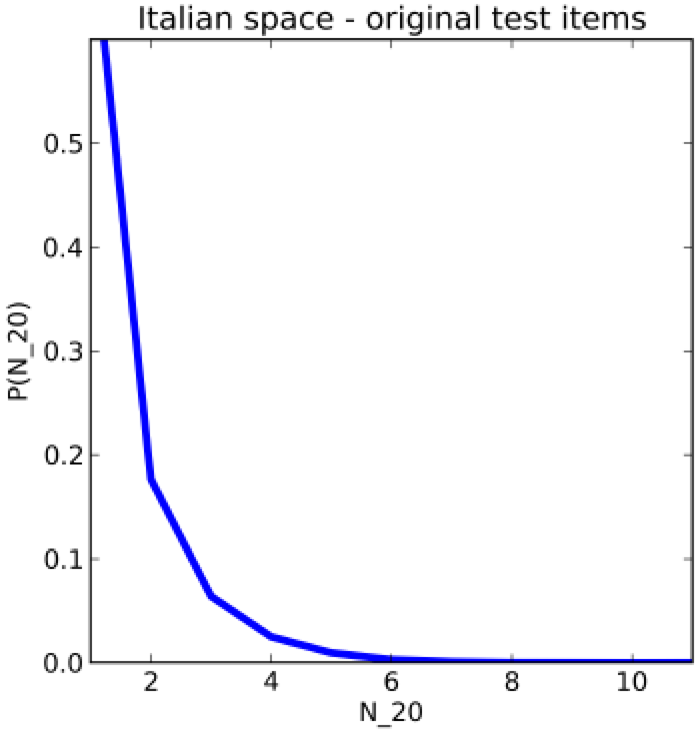
$$\max_{y \in Y} \cos(embed(y), y') \rightarrow \text{water}$$

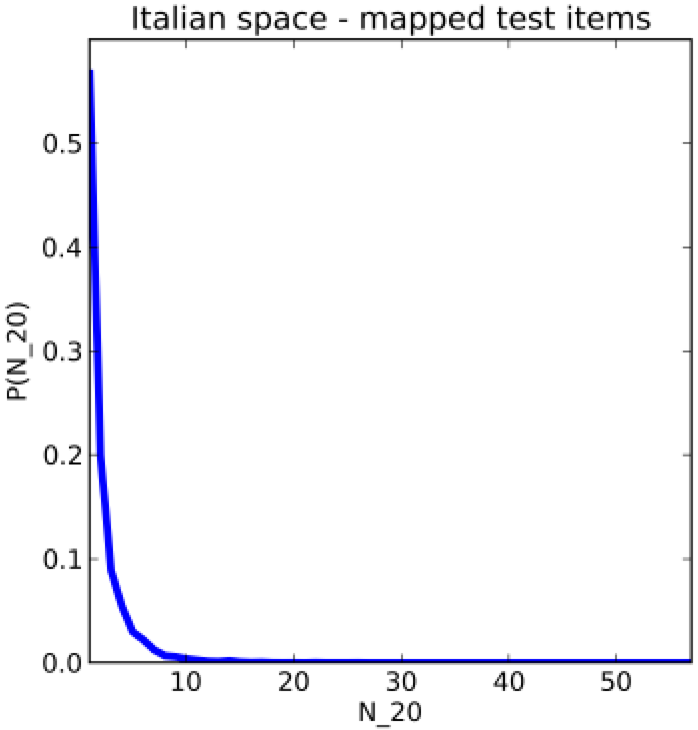
A standard intrinsic evaluation task for judging quality of cross-lingual embedding quality

The Hubness Problem with Nearest Neighbour

In high dimensional spaces, some points are neighbours of many points ➔ hubs

Adversely impacts Nearest Neigbour search ➔ especially in mapped spaces





Why does hubness occur?

- *Points are closer in mapped space with least-squares?*
- *Pairwise similarities tend to converge to constant as dimensionality increases*

Solutions to Hubness

Modify the search algorithm

- Inverted Rank (IR)
- Inverted Softmax (ISF)
- Cross-domain Similarity Local Scaling (CSLS)

Modify the learning objective to address hubness

- Max Margin Training
- Optimizing CSLS

Inverted Rank (Dinu et al., 2015)

$Rank_{a,Z}(z)$*: Rank of z in neighbourhood of a w.r.t candidate nodes Z*

In nearest neighbor we pick the target of rank 1

$$NN(x) = \operatorname*{argmin}_{y \in Y} Rank_{x,Y}(y)$$

In nearest neighbor we pick the target for which x has the lowest rank

$$IR(x) = \operatorname*{argmin}_{y \in Y} Rank_{y,X}(x)$$

Kind of collective classification, hubs will be assigned to the x to which they are closest

Inverted Softmax

(Smith et al., 2017)

Another way of inverse information lookup like IR

NN

$$P(y|x) = \frac{e^{\beta \cos(x,y)}}{\sum_{y'} e^{\beta \cos(x,y')}}$$

Distance Metric is generally normalized over target

ISF

$$P(y|x) = \frac{e^{\beta \cos(x,y)}}{\alpha_y \sum_{y'} e^{\beta \cos(x',y)}}$$

Modified Distance Metric normalized over source

Will penalize hubs since they have a large denominator

Local scaling of the distance metric

Cross-domain Similarity Local Scaling (CSLS)

(Conneau et al., 2018)

Another Local scaling of the distance metric

Define mean similarity of a mapped source word to its target neighbourhood and vice versa

$$r_T(x) = \frac{1}{K} \sum_{y \in N_T(x)} \cos(x, y) \qquad r_S(y) = \frac{1}{K} \sum_{x \in N_S(y)} \cos(x, y)$$

$$\boldsymbol{CSLS(x, y) = 2\cos(x, y) - r_T(x) - r_S(y)}$$

Will penalize hubs since they have large mean similarity

Symmetric metric
No parameter tuning

Optimizing CSLS

(Joulin et al., 2018)

For CSLS retrieval,

Training Metric: Cosine similarity *Test Metric: CSLS*

Mismatch between train and test metric

A good principle is to optimize for the objective we are interested in ➔ *optimize CSLS loss directly*

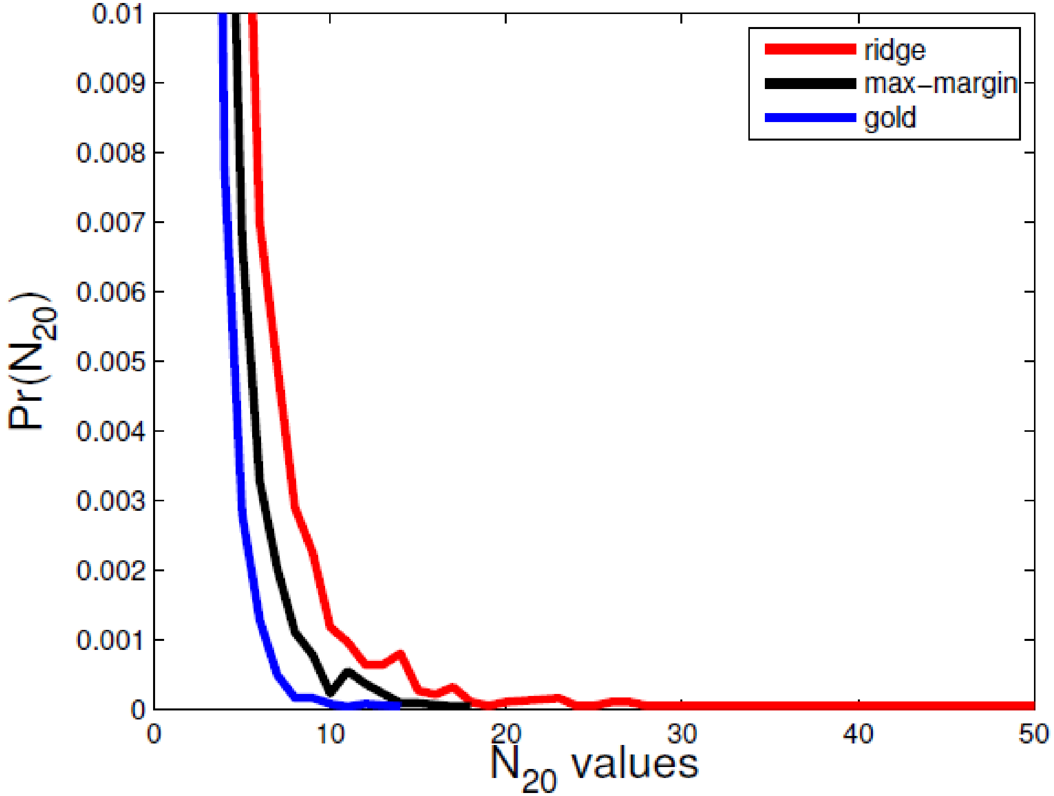
$$\boldsymbol{CSLS_{loss}(x,y) = -2\,cos(x,y) + r_T(x) + r_S(y)}$$

Max-Margin Formulation

(Lazaridou et al., 2015)

$$\sum_{j\neq i}^{N} \max\left\{0, \gamma + \|Wx_i - y_i\|^2 - \left\|Wx_i - y_j\right\|^2\right\}$$

Negative example must be as far good example as possible



Why would max-margin reduce hubness? ➔ No clear answer

Cross-Lingual Embeddings

Offline Methods

Online Methods (Slides adapted from Khapra and Chandar, 2016)

Some observations

Evaluation

Unsupervised Learning

Using Parallel Corpus Only

(Hermann and Blunsom, 2014)

Training data: Parallel sentences

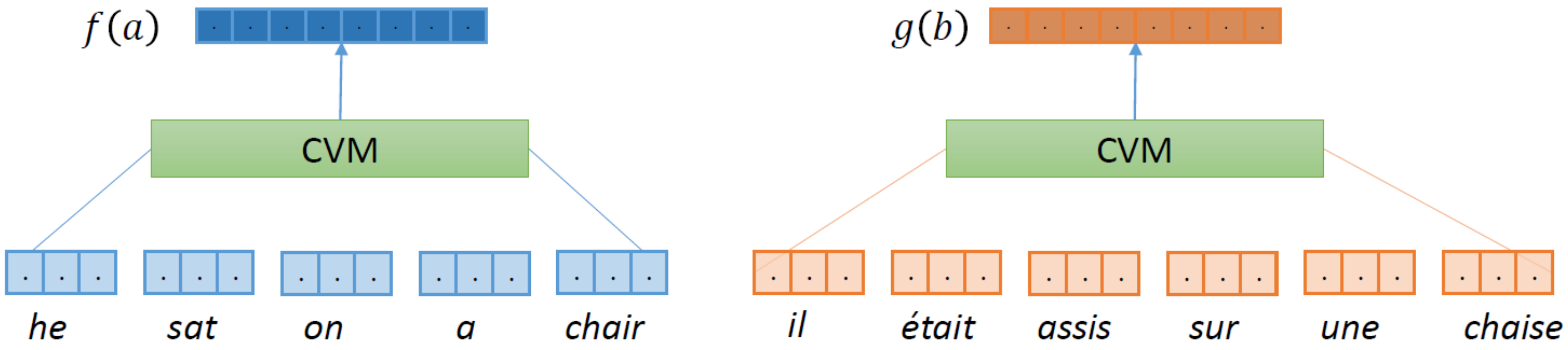
$a = English\ sentence$
$b = parallel\ French\ sentence$
$n = random\ French\ sentence$

$$E(a,b) = ||f(a) - g(b)||^2$$

$minimize$
$max(0, m + E(a,b) - E(a,n))$

Backpropagate & update w_i's in both languages

To reduce the distance between $f(a)$ & $g(b)$ the model will eventually learn to reduce the distance between (chair, chaise), (sit, assis), (he, il) etc.

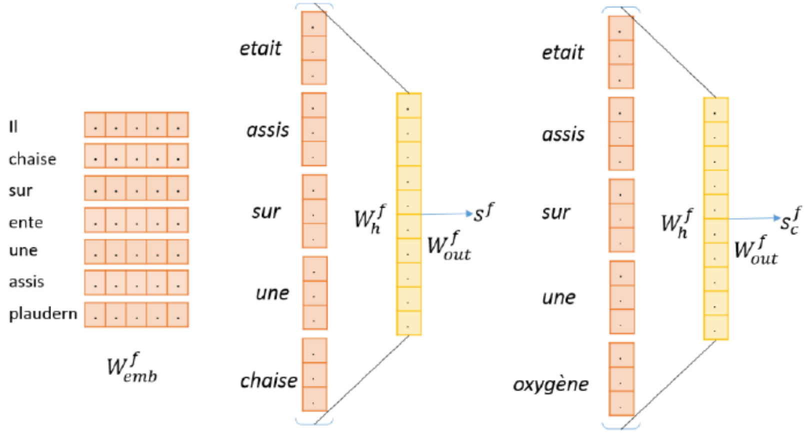


Using Parallel Corpus and Monolingual Corpus

(Gouws et al., 2015)

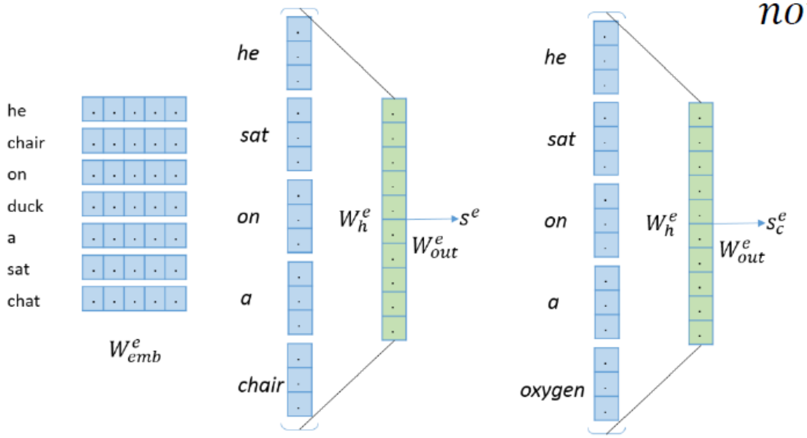
Fr positive: Il était assis sur une chaise

Fr negative: Il était assis sur une oxygène



En positive: he sat on a chair

En negative: he sat on a oxygen



Independently update θ^e and θ^f

$$\boxed{\begin{array}{l} maximize\ max(0, 1 - s^f + s^f_c) \\ w.r.t.\ \theta^e \end{array}}$$

+ Parallel data

En: he sat on a chair $[s_e = w^e_1, w^e_2, w^e_3, w^e_4, w^e_5]$

Fr : Il était assis sur une chaise $[s_f = w^f_1, w^f_2, w^f_3, w^f_4, w^f_5]$

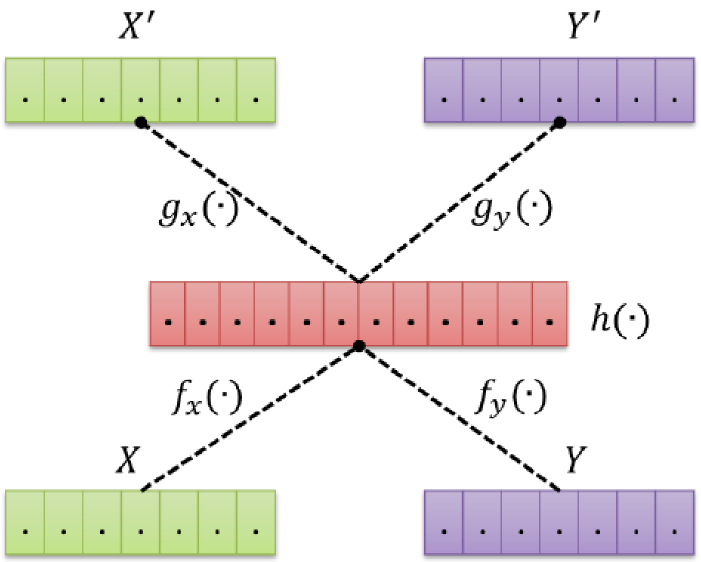
$$now, also\ minimize \quad \Omega\left(W^e_{emb}, W^f_{emb}\right) = \left\| \frac{1}{m} \sum_{w_i \in s^e}^{w_m} W^e_{emb_i} - \frac{1}{n} \sum_{w_j \in s^e}^{w_n} W^f_{emb_i} \right\|^2$$

$$w.r.t\ W^e_{emb}, W^f_{emb}$$

$$\boxed{\begin{array}{l} maximize\ max(0, 1 - s^e + s^e_c) \\ w.r.t.\ \theta^f \end{array}}$$

(Gouws et. al., 2015)

Using Parallel Corpus and Monolingual Corpus (Chandar et al., 2014)



A multiview autoencoder

N

encoder

$$h_x(X) = f_x(X) = f_x(\boldsymbol{W_x} X + b)$$

$$h_y(Y) = f_y(Y) = f_y(\boldsymbol{W_y} Y + b)$$

decoder

$$X' = g_x(h(X)) = g_x(\boldsymbol{W'_x} h_x(X) + b')$$

$$Y' = g_y(h(Y)) = g_y(\boldsymbol{W'_y} h_y(Y) + b')$$

$$minimize \sum_{i=1}^{N} (g_x(f_x(X_i)) - X_i)^2$$

$$+ \sum_{i=1}^{N} (g_y(f_y(Y_i)) - Y_i)^2$$

$$+ \sum_{i=1}^{N} (g_x(f_y(Y_i)) - X_i)^2$$

$$+ \sum_{i=1}^{N} (g_y(f_x(X_i)) - Y_i)^2$$

$$\boxed{-corr(h(\bar{X}), h(\bar{Y}))}$$

- *Autoencoder approach*
- *Correlation term is important to ensure common representation*
- *Combines:*
  - *word similarity (recall Procrustes!)*
  - *dimension correlation (recall CCA!)*

A general framework for cross-lingual embeddings

$$maximize \quad \sum_{j\in\{e,f\}} \sum_{i=1}^{T_j} \underbrace{-\log(P(w_i|w_{i-k},\dots,w_{i-1}))}_{\textit{monolingual similarity}} + \underbrace{\lambda\cdot\ \Omega\,(W^e_{emb}, W^f_{emb})}_{\textit{bilingual similarity}}$$

$$w.r.t\ \ \theta_e, \theta_f$$

$$\theta_e = W^e_{emb}, W^e_h, W^e_{out}$$

$$\theta_f = W^f_{emb}, W^f_h, W^f_{out}$$

$$\Omega\left(W^e_{emb}, W^f_{emb}\right) = \sum_{w_i\in V^e} \sum_{w_j\in V^f} sim\left(w_i, w_j\right) * distance(W^e_{emb_i}, W^f_{emb_j})$$

This weighted sum will be low only when similar words across languages are embedded close to each other

Offline embeddings also follow this framework, but they optimize the monolingual and bilingual objectives sequentially

Cross-Lingual Embeddings

Offline Methods

Online Methods

Some observations

Evaluation

Unsupervised Learning

Intrinsic Evaluation

- Bilingual Lexicon Induction

- Cross-language word similarity task

Mostly offline methods

Bilingual Lexicon Induction

	English to Italian			Italian to English		
	P@1	**P@5**	**P@10**	**P@1**	**P@5**	**P@10**
Ordinary Least Squares	33.8	48.3	53.9	24.9	41.0	47.4
OP + NN	36.9	52.7	57.9	32.2	49.6	55.7
OP + IR	38.5	56.4	63.9	24.6	45.4	54.1
OP + ISF	43.1	60.7	66.4	38.0	**58.5**	**63.6**
OP + CSLS	44.9	**61.8**	**66.6**	**38.5**	57.2	63.0
OP + CSLS (optimize)	**45.3**	NA	NA	37.9	NA	NA
CCA	36.1	52.7	58.1	31.0	49.9	57.0

Orthogonality constraint helps

Bilingual Lexicon Induction

	English to Italian			Italian to English		
	P@1	**P@5**	**P@10**	**P@1**	**P@5**	**P@10**
Ordinary Least Squares	33.8	48.3	53.9	24.9	41.0	47.4
OP + NN	36.9	52.7	57.9	32.2	49.6	55.7
OP + IR	38.5	56.4	63.9	24.6	45.4	54.1
OP + ISF	43.1	60.7	66.4	38.0	**58.5**	**63.6**
OP + CSLS	44.9	**61.8**	**66.6**	**38.5**	57.2	63.0
OP + CSLS (optimize)	**45.3**	NA	NA	37.9	NA	NA
CCA	36.1	52.7	58.1	31.0	49.9	57.0

Modified retrieval significantly improve performance over vanilla Nearest Neighbour Search

CSLS is best performing

Optimizing CSLS loss also gives some improvements

Bilingual Lexicon Induction

	English to Italian			Italian to English		
	P@1	**P@5**	**P@10**	**P@1**	**P@5**	**P@10**
Ordinary Least Squares	33.8	48.3	53.9	24.9	41.0	47.4
OP + NN	36.9	52.7	57.9	32.2	49.6	55.7
OP + IR	38.5	56.4	63.9	24.6	45.4	54.1
OP + ISF	43.1	60.7	66.4	38.0	**58.5**	**63.6**
OP + CSLS	44.9	**61.8**	**66.6**	**38.5**	57.2	63.0
OP + CSLS (optimize)	**45.3**	NA	NA	37.9	NA	NA
CCA	36.1	52.7	58.1	31.0	49.9	57.0

Orthogonal Procrustes solution and CCA give roughly the same results

Extrinsic Evaluation

- Cross-lingual Document Classification
- Cross-lingual Dependency Parsing

Mostly online methods

Cross-lingual Document Classification

Approach	en→ de	de → en
Hermann & Blunson, 2014	83.7	71.4
Chandar et al., 2014	91.8	72.8
Gouws et al., 2015	86.5	75.0

Leveraging monolingual and parallel corpora yields better results

Cross-Lingual Embeddings

Offline Methods

Online Methods

Some observations

Evaluation

Unsupervised Learning

More observations on different aspects of the problem

Take them with a pinch of salt, since comprehensive experimentation is lacking

More like rule of thumb to make decisions

Effect of bilingual dictionary size

(Dinu et al., 2015)

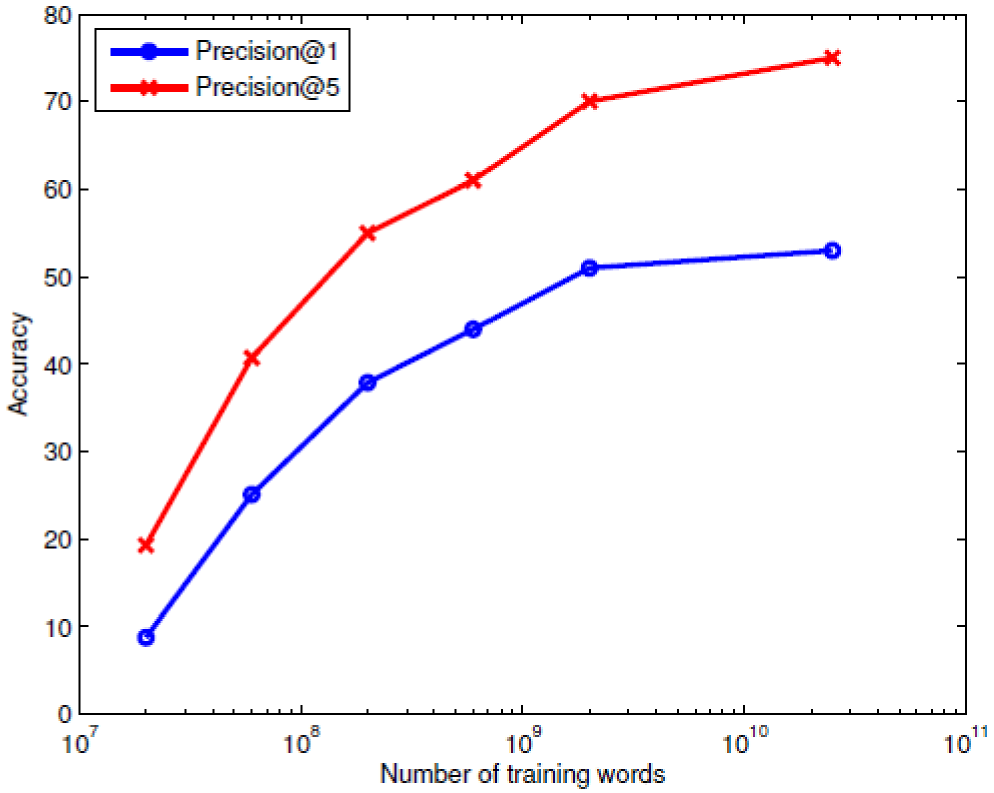
Dictionary Size	Precision@1
1K	20.09
5K	37.3
10K	37.5
20K	37.9

Beyond a certain size, the size of bilingual dictionary does not seem useful

What if the bilingual dictionaries are really large?

Effect of monolingual corpora size

(Mikolov et al., 2013)

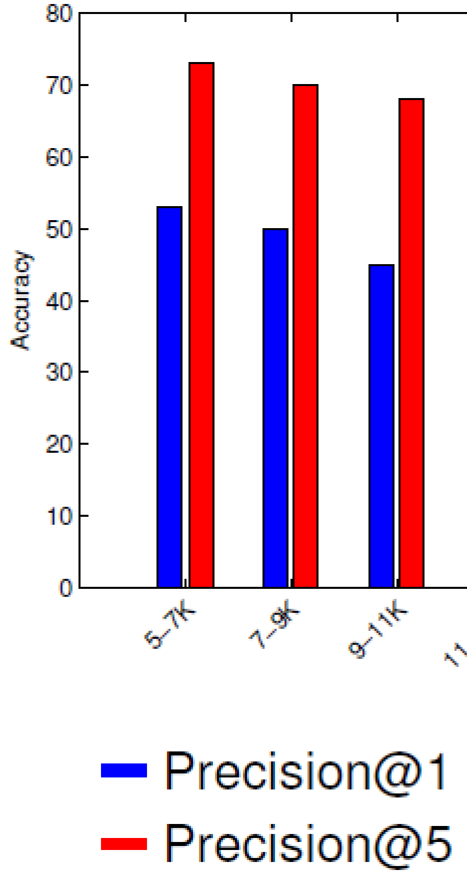


Large monolingual corpora substantially increases the quality of embeddings

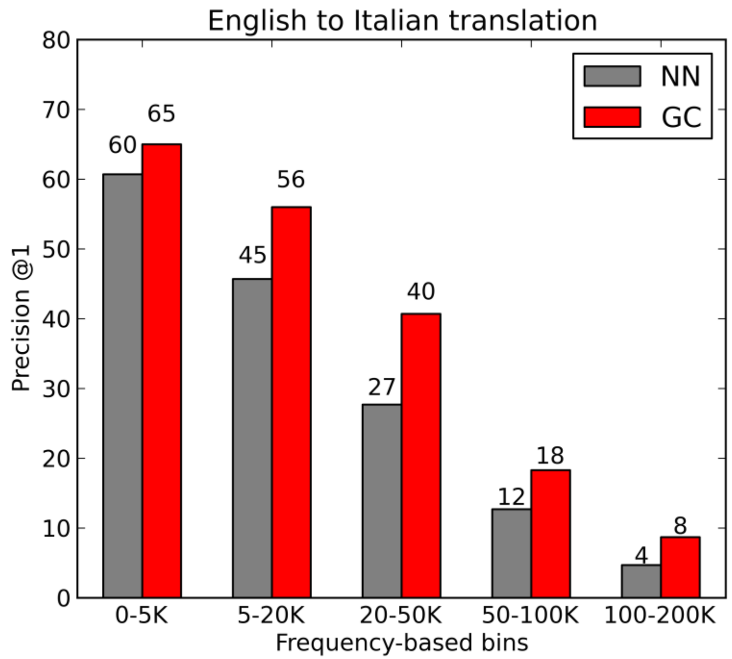
Having large monolingual corpora may be more useful than having large bilingual dictionary?

How difficult is to translate less frequent words?

- Performance does not drop very sharply for intermediate frequency words
- Performance drops sharply for very rare words



(Mikolov et al., 2013)



(Dinu et al., 2015)

Note: GC is same as Inverse Rank retrieval

Do these approaches work for all languages?

https://github.com/Babylonpartners/fastText_multilingual#right-now-prove-that-this-procedure-actually-worked

- *Study on 78 languages*
- *Trained on 10k words (Dictionary created using Google Translate)*
- *Tested on 2500 words*
- *Method described by Smith et al., 2017 (Procrustes with inverted softmax)*

Best Languages	Worst Languages
French	Urdu
Portuguese	Marathi
Spanish	Japanese
Norwegian	Punjabi
Dutch	Burmese
Czech	Luxembourgish
Hungarian	Malagasy

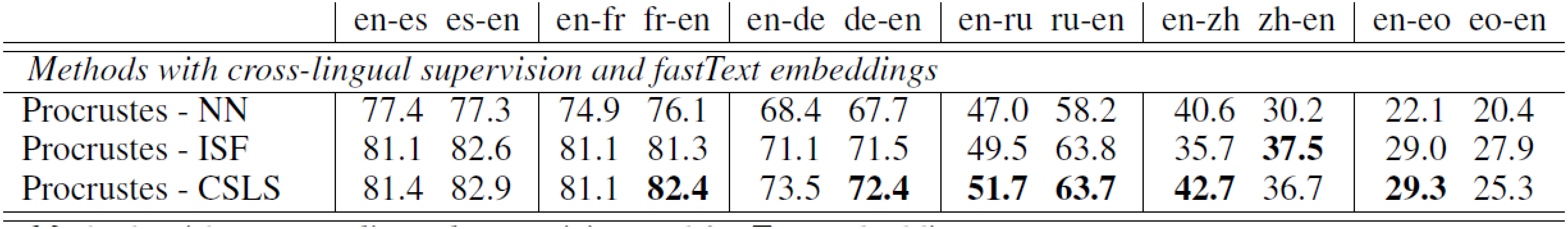
No patterns, seems to be a function of dictionary quality in each language

Facebook has recently provided high quality bilingual dictionaries ➔ a testbed to do better testing

https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

Do these approaches work for all languages?

Results on more languages from Conneau et al., 2018



	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
Methods with cross-lingual supervision and fastText embeddings												
Procrustes - NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	22.1	20.4
Procrustes - ISF	81.1	82.6	81.1	81.3	71.1	71.5	49.5	63.8	35.7	**37.5**	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	**82.4**	73.5	**72.4**	**51.7**	**63.7**	**42.7**	36.7	**29.3**	25.3

Seems to work well on mainland European languages compared to Russian, Chinese and Esperanto

Cross-Lingual Embeddings

Offline Methods

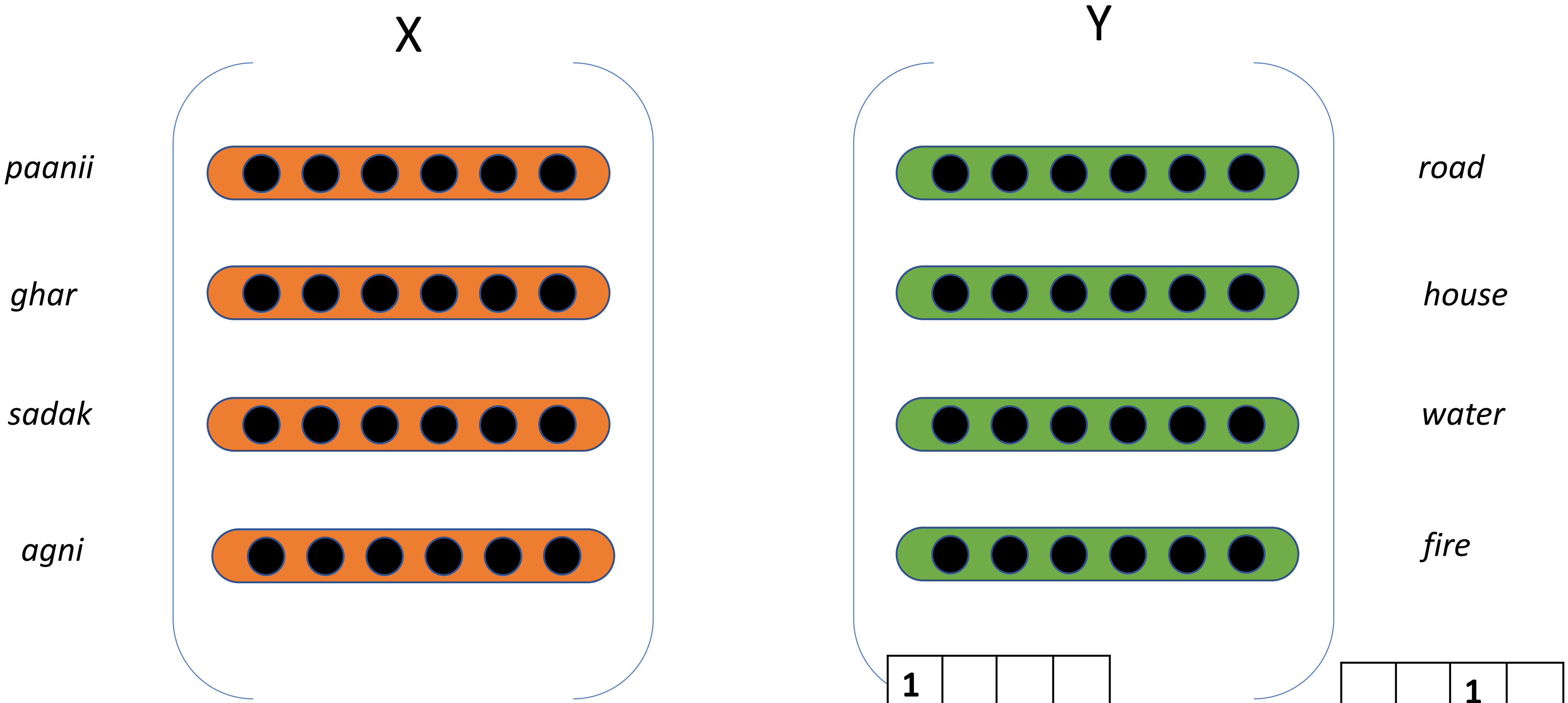
Online Methods

Some observations

Evaluation

Unsupervised Learning

Unsupervised Learning



$$\mathrm{X}\mathrm{W} = PY$$

$$P = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \rightarrow \begin{bmatrix} & & 1 & \\ & 1 & & \\ 1 & & & \\ & & & 1 \end{bmatrix}$$

(Permutation matrix)

Many language pairs may not have an available bilingual dictionary

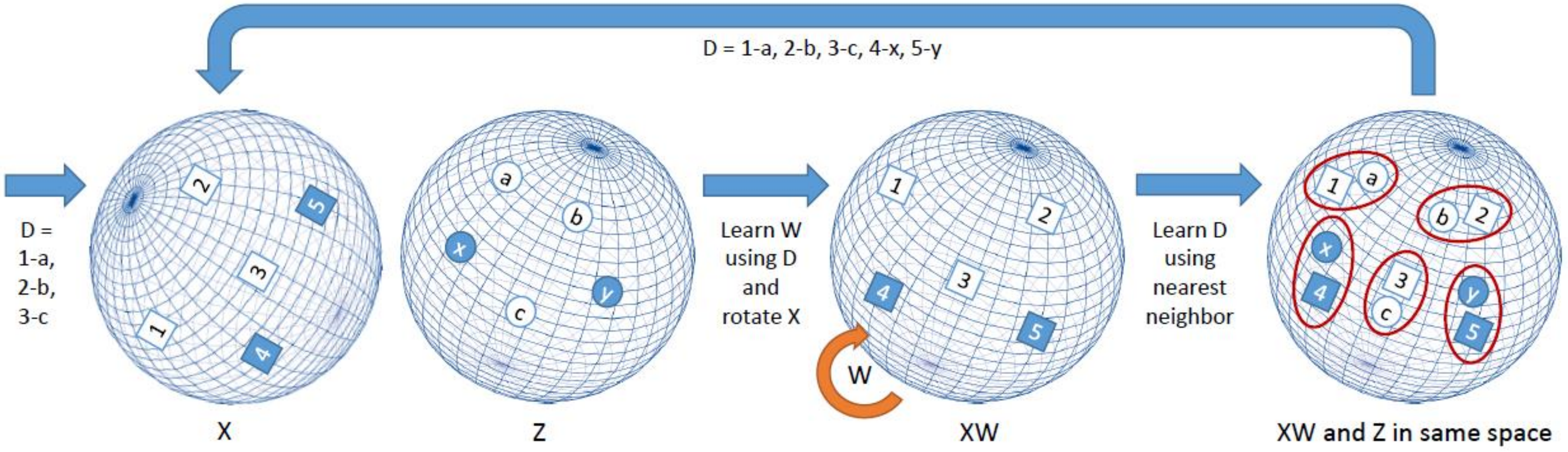
Mostly offline methods – by definition

Exciting developments on this task this year

Starting with a small seed dictionary

(Artetxe et al., 2017)

- As small as 50-100
- Dictionary can just be aligned digits and numbers
  - १ → 1
  - २८९ → 289
  - ५ → 5
- Identical strings
  - Requires both languages to have similar scripts and share vocabulary
- Bootstrapping solution



$$W^* = \arg\max_W \sum_i \max_j \left(X_{i*}W\right) \cdot Z_{j*}$$

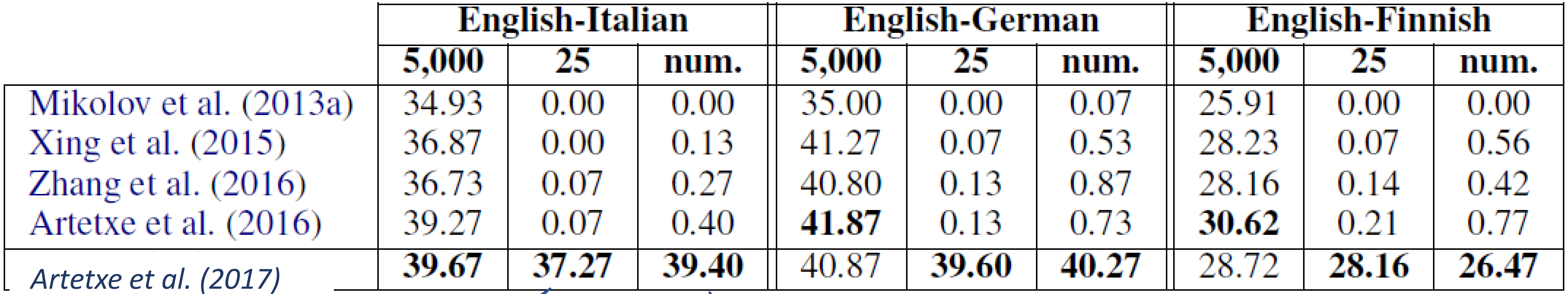
$$\text{s.t.} \quad WW^T = W^TW = I$$

Enhancements by Hoshen and Wolf (2018)

- *do away with the need for seed dictionary by matching principal components for initialization*
- *consider a objective in other direction and circular objective too*

Enhancements by Artetxe et al., (2018b)

- *do away with the need for seed dictionary by using word similarity distribution for initialization*



	English-Italian			English-German			English-Finnish		
	5,000	**25**	**num.**	**5,000**	**25**	**num.**	**5,000**	**25**	**num.**
Mikolov et al. (2013a)	34.93	0.00	0.00	35.00	0.00	0.07	25.91	0.00	0.00
Xing et al. (2015)	36.87	0.00	0.13	41.27	0.07	0.53	28.23	0.07	0.56
Zhang et al. (2016)	36.73	0.07	0.27	40.80	0.13	0.87	28.16	0.14	0.42
Artetxe et al. (2016)	39.27	0.07	0.40	**41.87**	0.13	0.73	**30.62**	0.21	0.77
Artetxe et al. (2017)	**39.67**	**37.27**	**39.40**	40.87	**39.60**	**40.27**	28.72	**28.16**	**26.47**

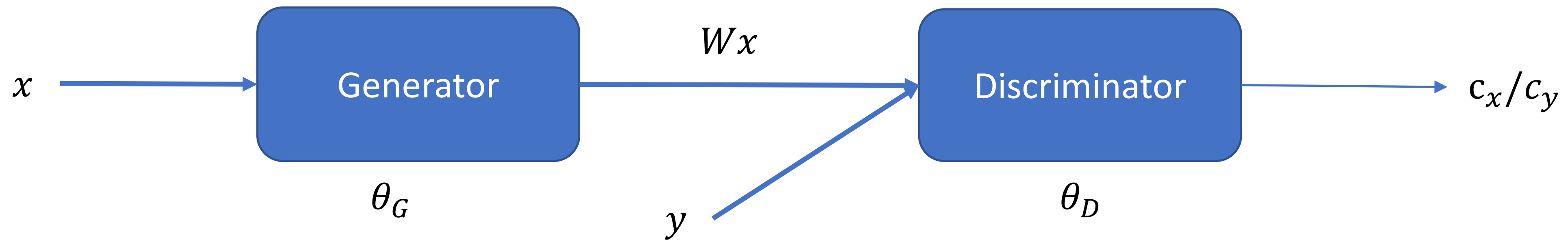
Source: Artetxe et al., (2017)

Bootstrapping works well with small dictionaries

Aligned numbers are sufficient to bootstrap

Adversarial Training

(Barone, 2016; Zhang et al., 2017a,b; Conneau et al., 2018)



We want to make Wx and y indistinguishable

Step 1: Make a good discriminator that can distinguish between Wx and y (optimize θ_D)

Step 2: Try to fool this discriminator by generating Wx which are indistinguishable (optimize θ_G)

Iterate with improved generator

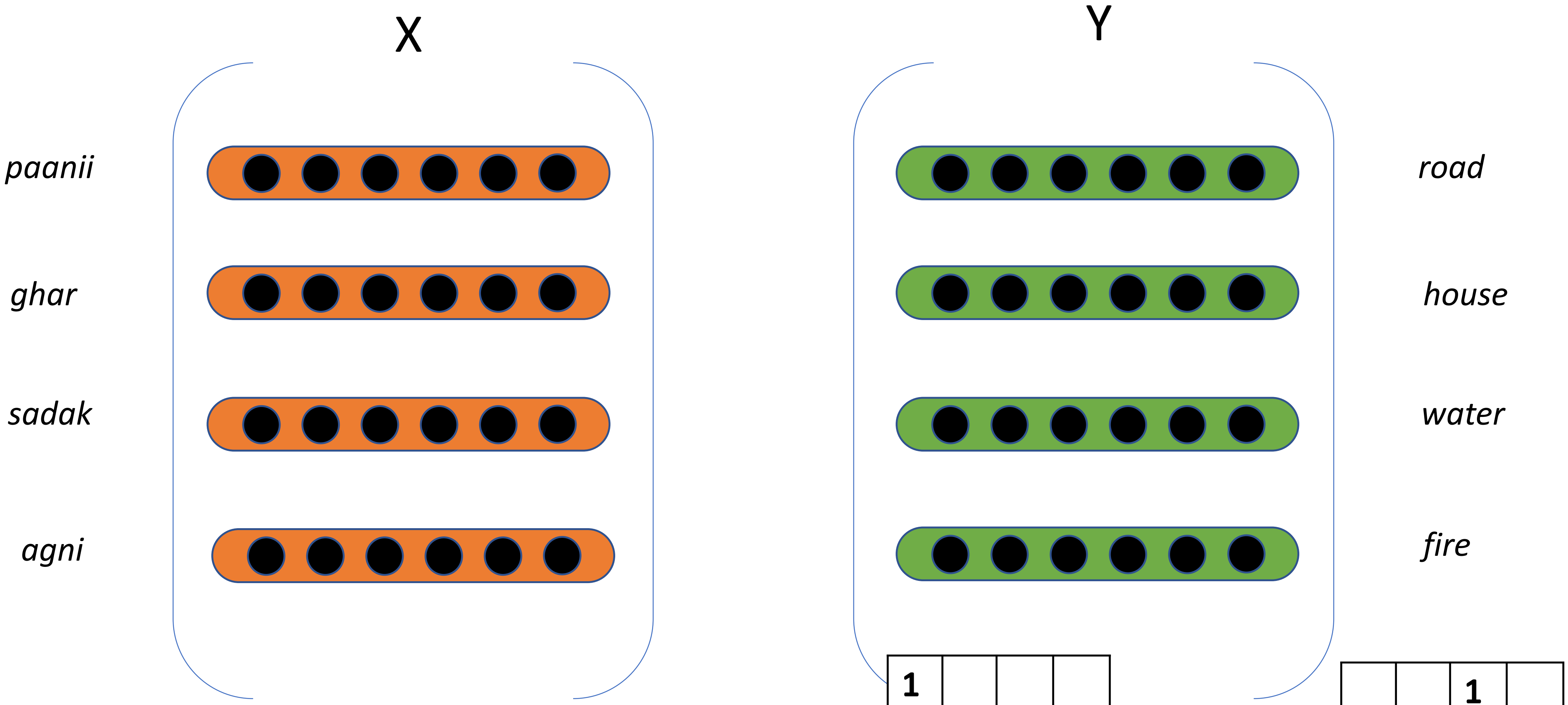
Conneau et al., 2018 suggested multiple runs, rebuilding & refining dictionary after each run

Tips for training

- Training adversarial networks is not easy – have to balance two objectives
- There may be a mismatch between discriminator and task classifier quality
- *e.g* If the discriminator is weaker
    - Design training schedule s.t. early epochs focus on improving the classifier
- Stabilizing GAN training is an active area of work

Wasserstein Procrustes

(Zhang et al., 2017b; Grave et al., 2018)



$$XW = PY$$

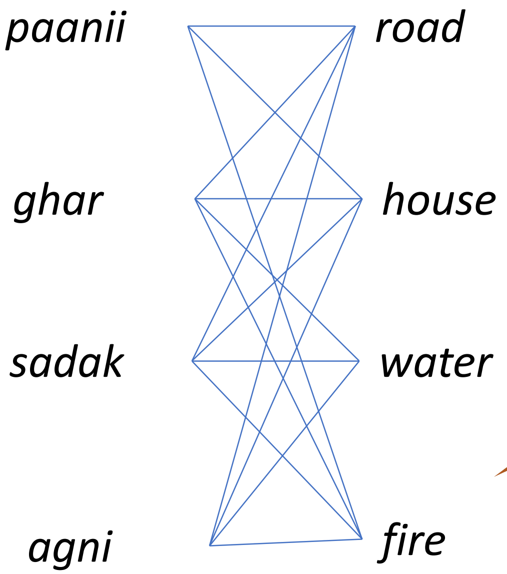
$$P = \begin{bmatrix}1&&&\\&1&&\\&&1&\\&&&1\end{bmatrix} \rightarrow \begin{bmatrix}&&1&\\&1&&\\1&&&\\&&&1\end{bmatrix}$$

(Permutation matrix)

If P is known, we can find W using the orthogonal Procrustes solution

$$W^* = \underset{W \in O_d}{\operatorname{argmin}} \|XW - PY\|_2^2$$

If W is known, finding P is equivalent to finding maximum weight matching in a bipartite graph



Solution

Hungarian Algorithm

equivalent to

Wasserstein Distance

$$P^* = \min_P \sum_{i,j} P_{ij} \left\| x_i W - y_j \right\|_2^2$$

Approximate solution using the Sinkhorn algorithm

Edge-weight(a,b) = - distance(a,b)

The dataset as a whole is aligned, considering constraints from all examples

Overall, problem is

$$\min_{W \in O_d} \min_{P} \|XW - PY\|_2^2$$

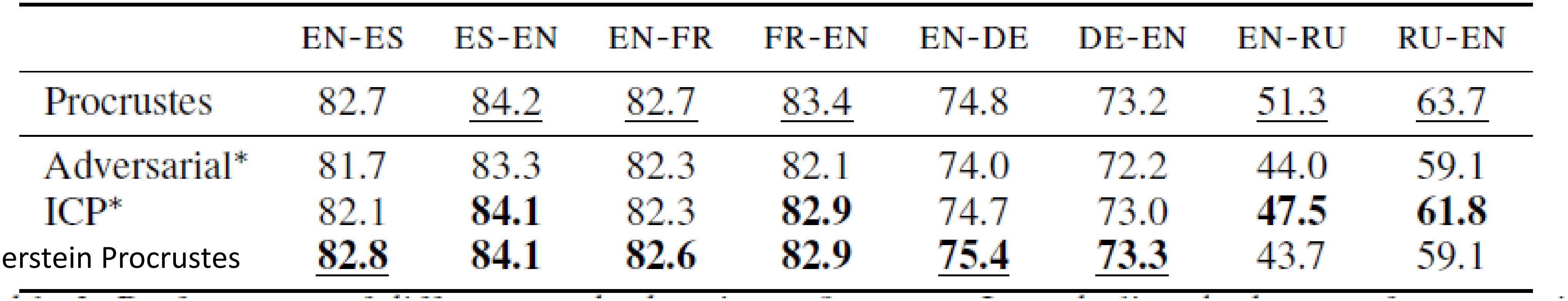
We can solve each minimization problem alternately, keep the other parameter constant

Good initialization of the problem is important

Grave et al., 2018 suggest a convex relaxation of the above problem

The solution to the convex relaxation is a good initializer to the problem

Comparing unsupervised methods



	EN-ES	ES-EN	EN-FR	FR-EN	EN-DE	DE-EN	EN-RU	RU-EN
Procrustes	82.7	84.2	82.7	83.4	74.8	73.2	51.3	63.7
Adversarial*	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1
ICP*	82.1	**84.1**	82.3	**82.9**	74.7	73.0	**47.5**	**61.8**
Wasserstein Procrustes	**82.8**	**84.1**	**82.6**	**82.9**	**75.4**	**73.3**	43.7	59.1

Source: Grave et al., (2018)

- Unsupervised methods can rival supervised approaches
- Even linear transformation based methods can perform well
- Shows the strong structural correspondence between embedding spaces across languages
- A launchpad for unsupervised sentence translation

Outline

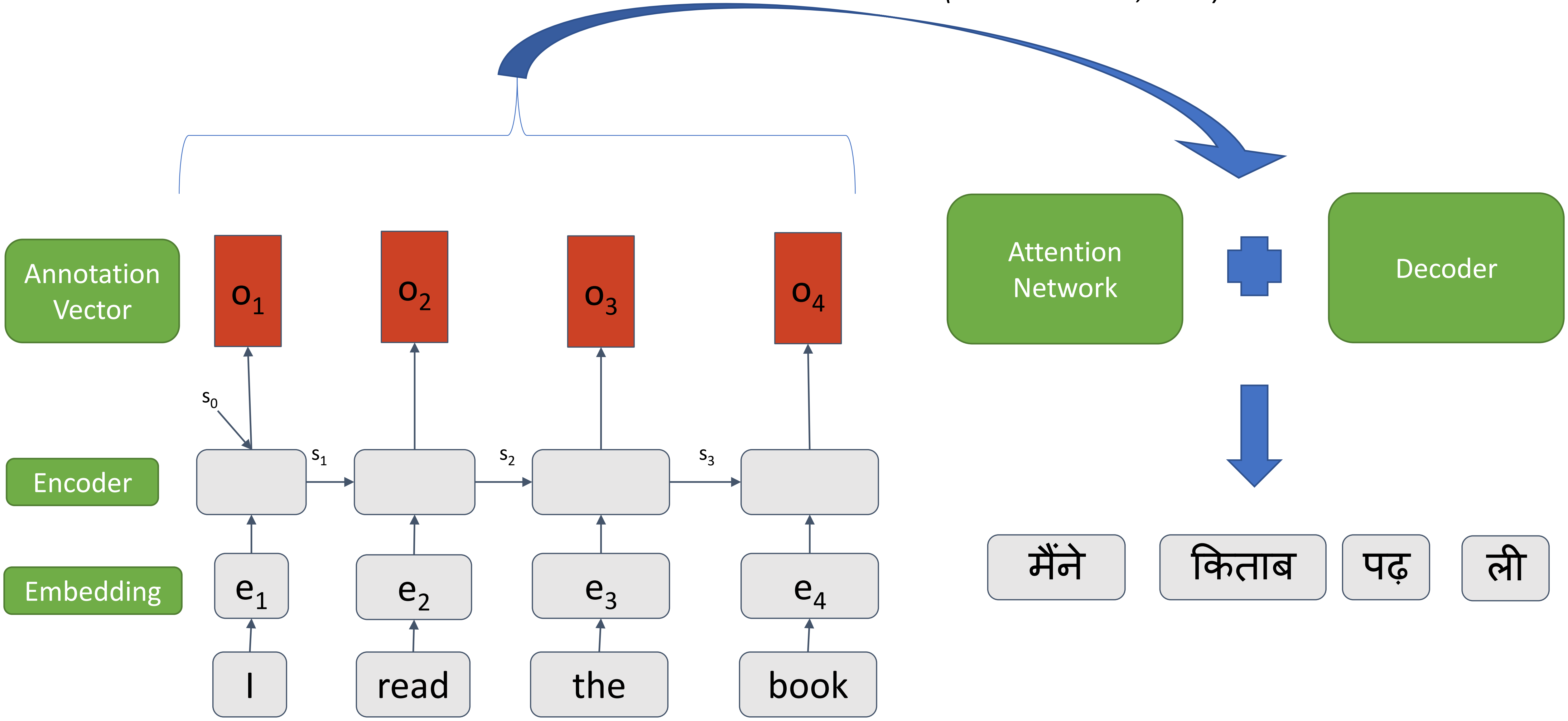
- Learning Cross-lingual Embeddings
- Training a Multilingual NLP Application
- Related Languages and Multilingual Learning
- Summary and Research Directions

Multilingual Neural Machine Translation

A Case Study

Embed - Encode - Attend - Decode Paradigm

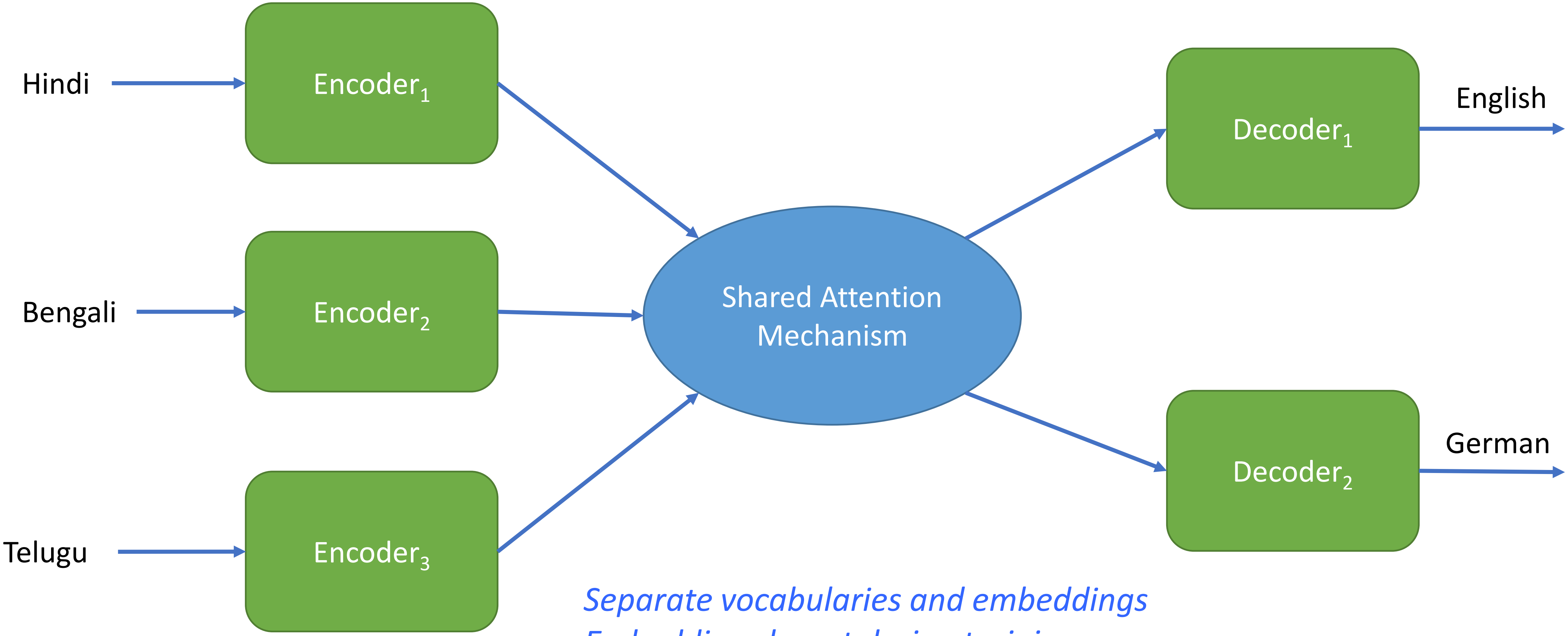
(Bahdanau et al, 2015)



Joint Learning

Minimal Parameter Sharing

(Firat et al., 2016)



Separate vocabularies and embeddings

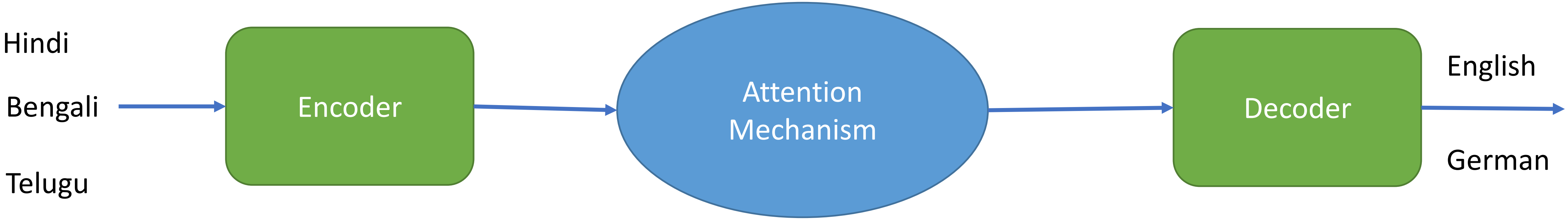
Embeddings learnt during training

Source Embeddings projected to a common space

Cycle through each language pair in minibatches

All Shared Architecture

(Johnson et al., 2017)



Shared vocabularies and embeddings across languages

Embeddings learnt during training

Source Embeddings projected to a common space

A minibatch contains data from all language pairs

How do we support multiple target languages with a single decoder?

A simple trick!

Append input with special token indicating the target language

For English-Hindi Translation

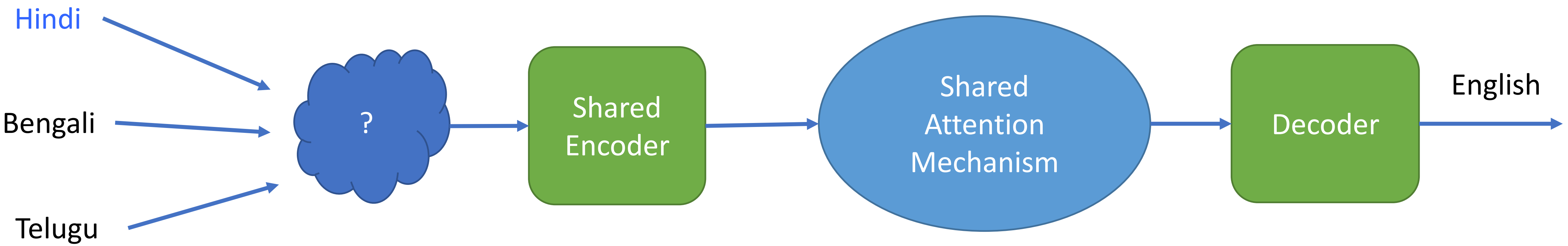
Original Input: *France and Croatia will play the final on Sunday*

Modified Input: *France and Croatia will play the final on Sunday* *<hin>*

Transfer Learning

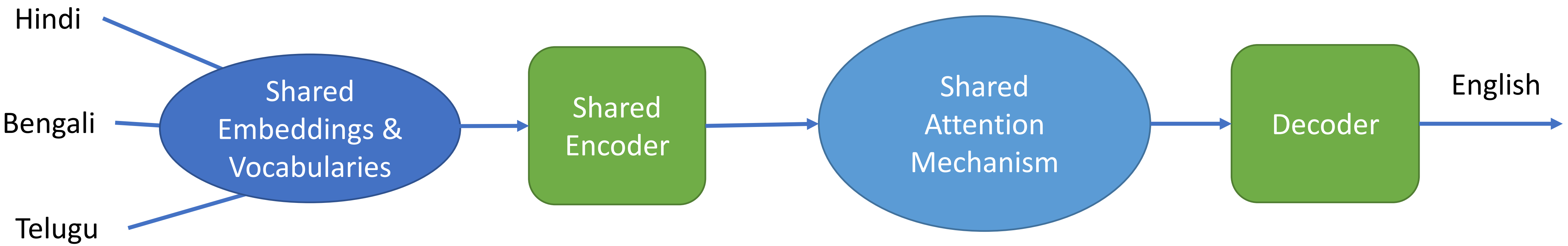
Shared Encoder

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017)



Shared Encoder

(Zoph et al., 2016; Nguyen and Chang, 2017; Lee et al., 2017)



Zoph et al., 2016: Randomly map primary and assisting language word embeddings

Lee et al., 2017: Character as basic unit

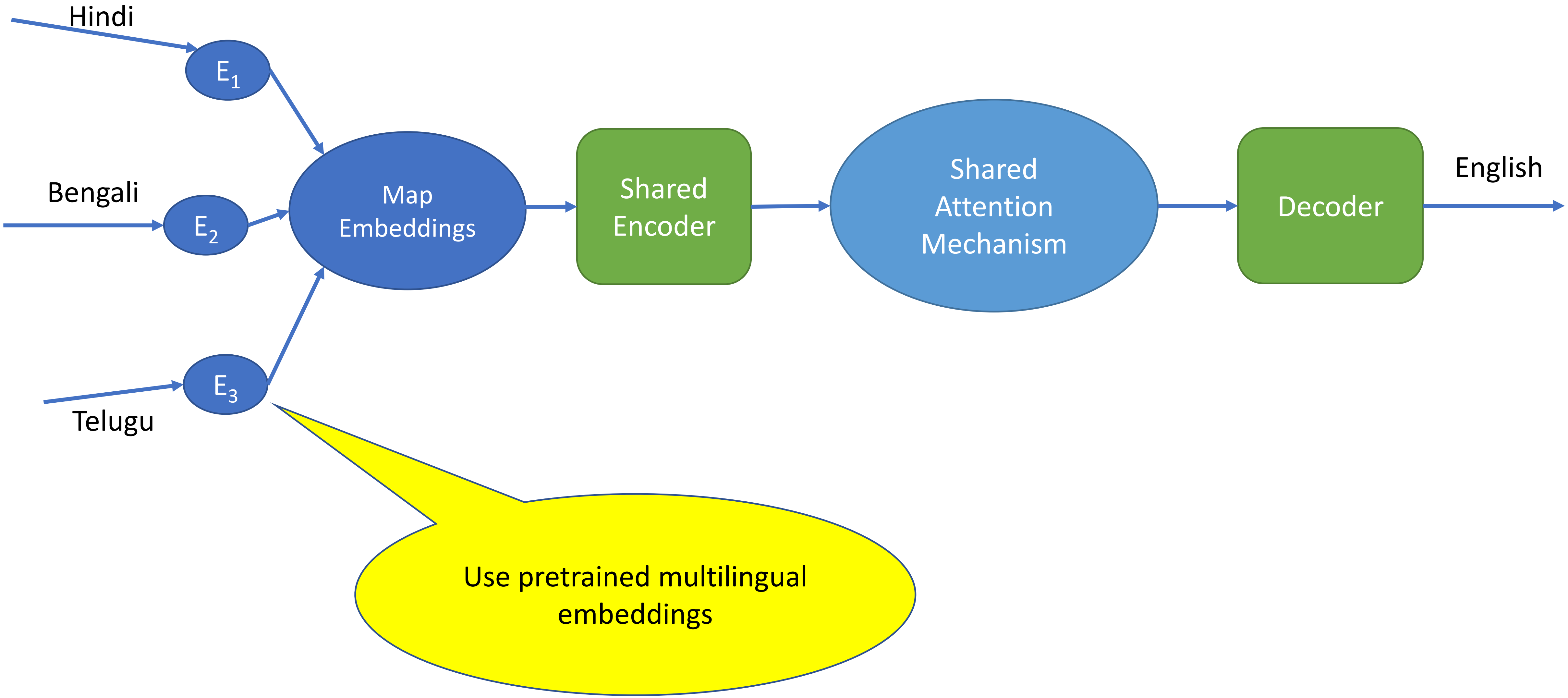
Single vocabulary as long as primary and assisting languages have compatible scripts

Nguyen et al., 2017: Use BPE to learn a common vocabulary across primary and assisting languages

BPE identifies small substring patterns in text

Shared Encoder

(Gu et al., 2018)

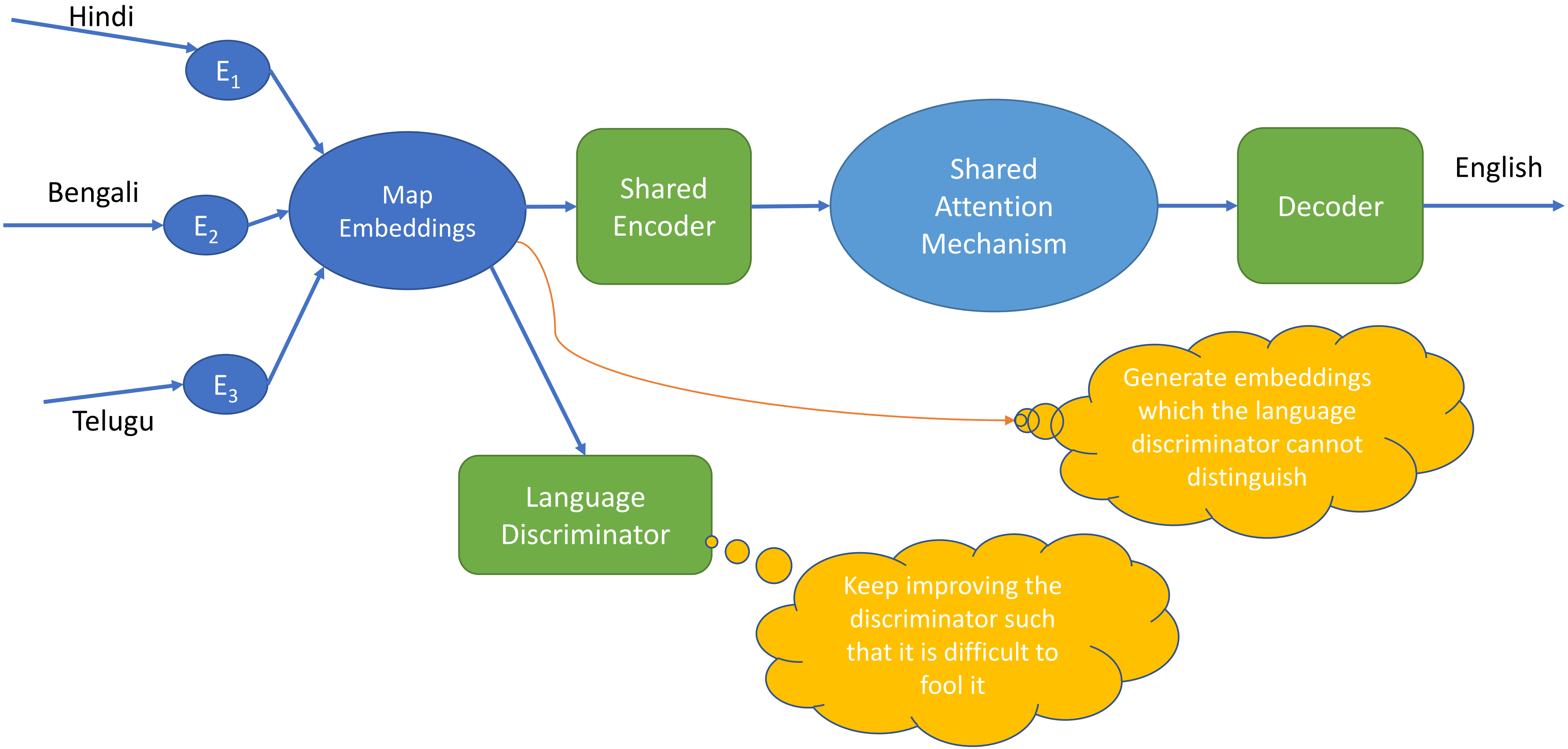


Use pretrained multilingual embeddings

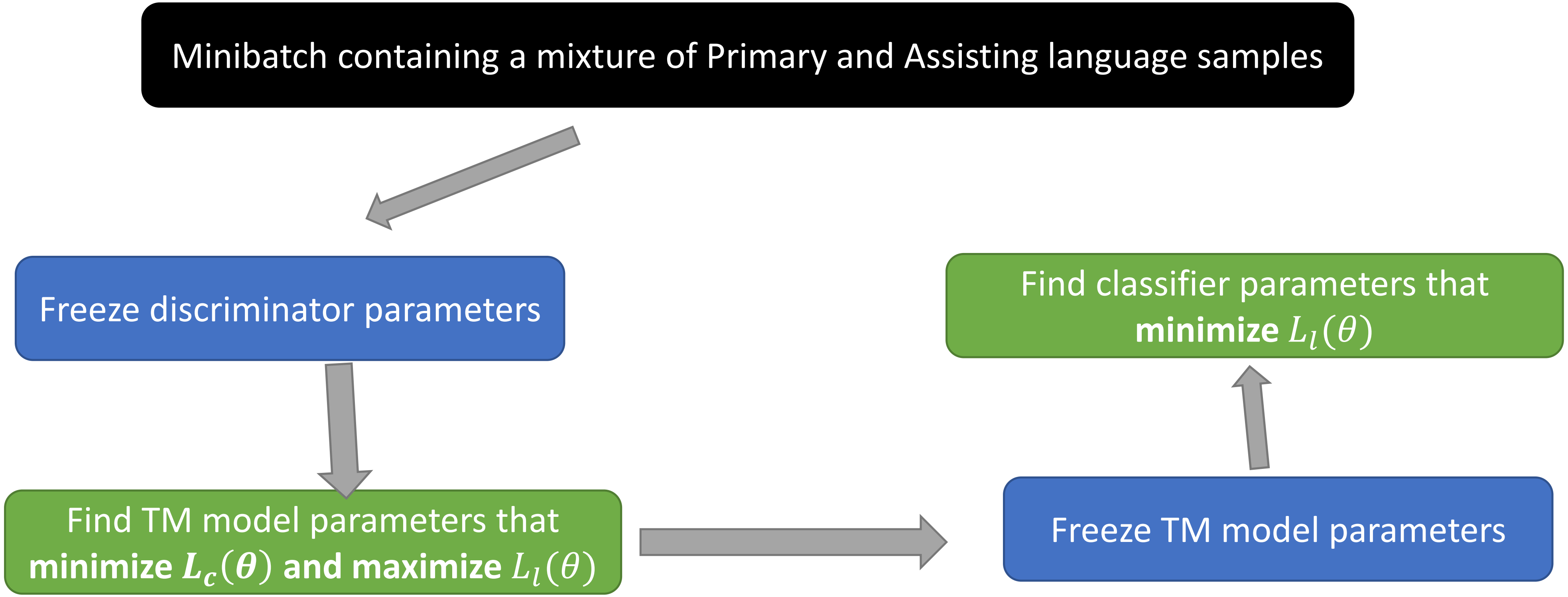
How do we ensure that encoder representations are similar across languages?

Shared Encoder with Adversarial Training

(Joty et al., 2017)


Hindi
Bengali
Telugu
E_1
E_2
E_3
Map Embeddings
Shared Encoder
Shared Attention Mechanism
Decoder
English
Language Discriminator
Generate embeddings which the language discriminator cannot distinguish
Keep improving the discriminator such that it is difficult to fool it

Training Process



Minibatch containing a mixture of Primary and Assisting language samples

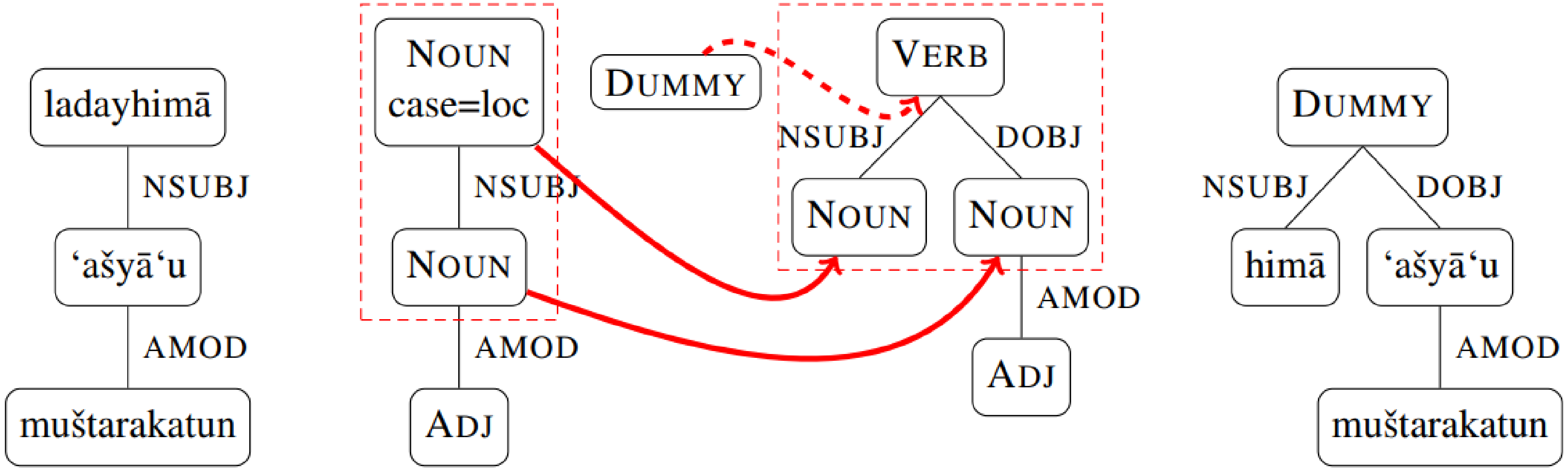
Freeze discriminator parameters

Find TM model parameters that **minimize $L_c(\theta)$ and maximize** $L_l(\theta)$

Freeze TM model parameters

Find classifier parameters that **minimize** $L_l(\theta)$

Preprocess Sentences *(Ponti et al., 2018)*



(a) Original (source) tree (b) Match templates (c) Processed (source) tree

Data Selection

(Rudramurthy et al., 2018)

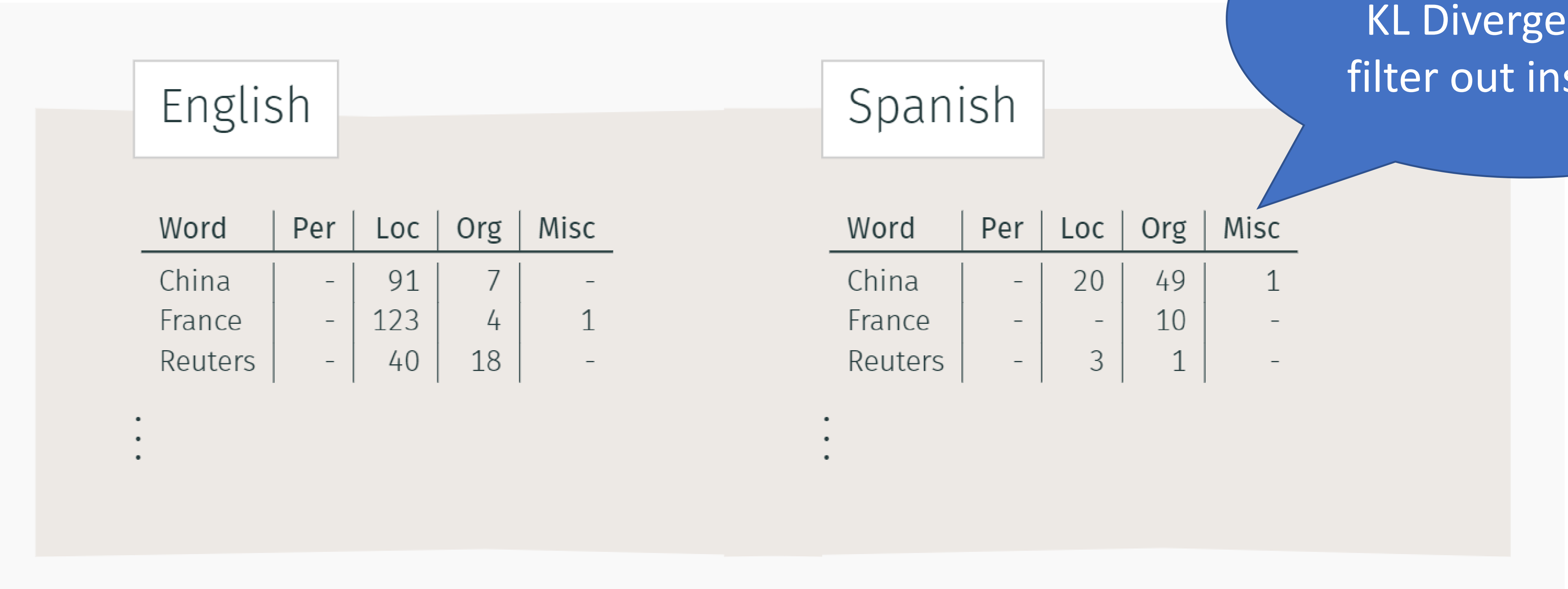
Is all the high-resource assisting language data useful?

Maybe, sentences with a very different structure from primary language are harmful

Let’s take a simpler example → Named Entity Recognition

Filter out training examples with high tag distribution divergence

Measure Symmetric KL Divergence to filter out instances



English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1
Reuters	-	40	18	-

⋮

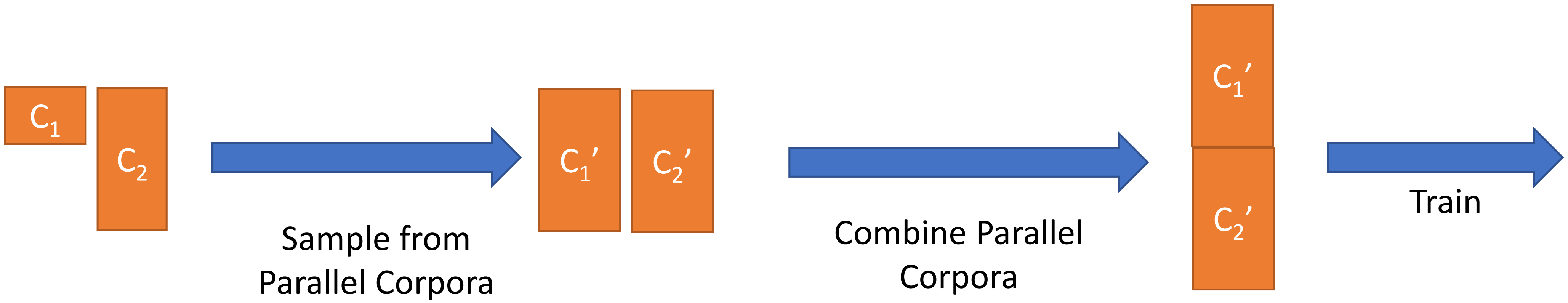
Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
France	-	-	10	-
Reuters	-	3	1	-

⋮

Training Transfer learning systems

Method 1



C_1 C_2

Sample from Parallel Corpora

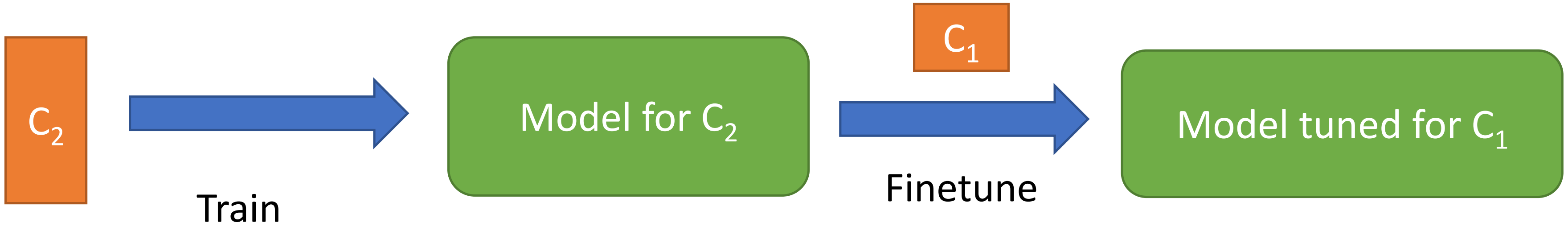
C_1' C_2'

Combine Parallel Corpora

C_1' C_2'

Train

Method 2



C_2

Train

Model for C_2

C_1

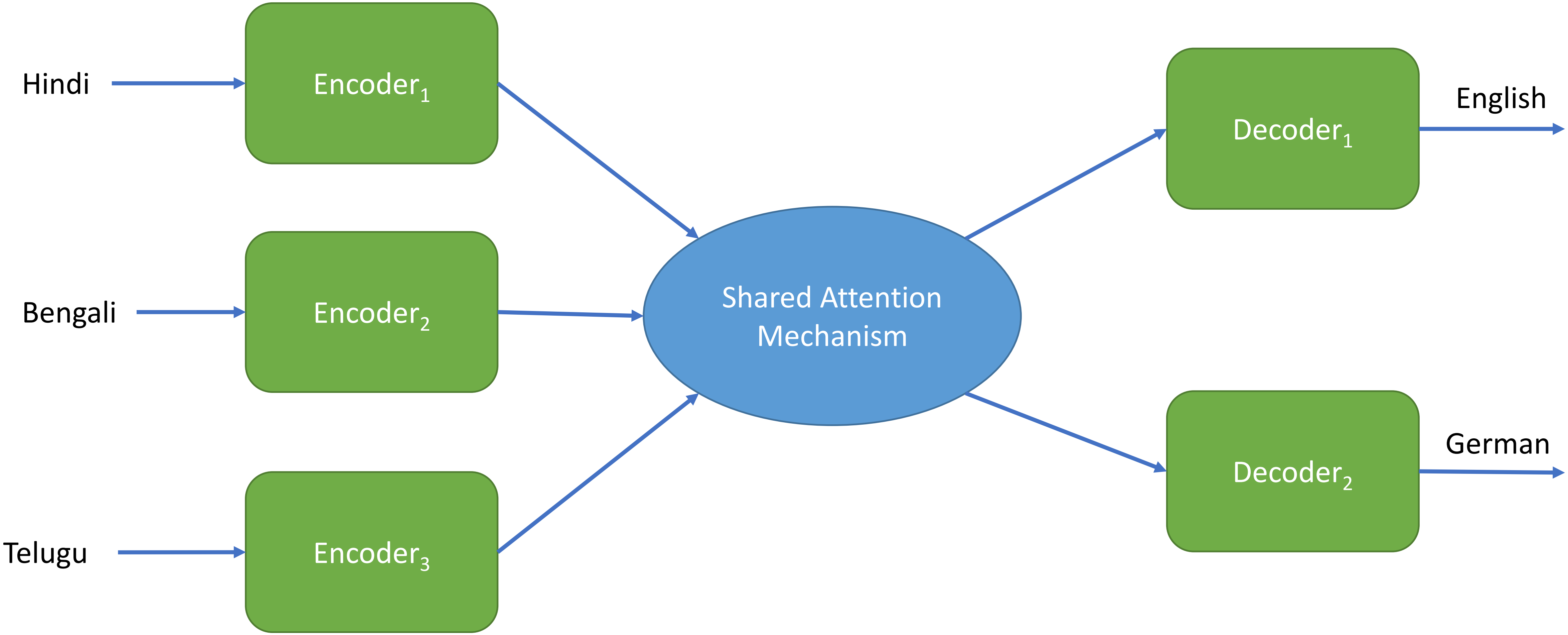
Finetune

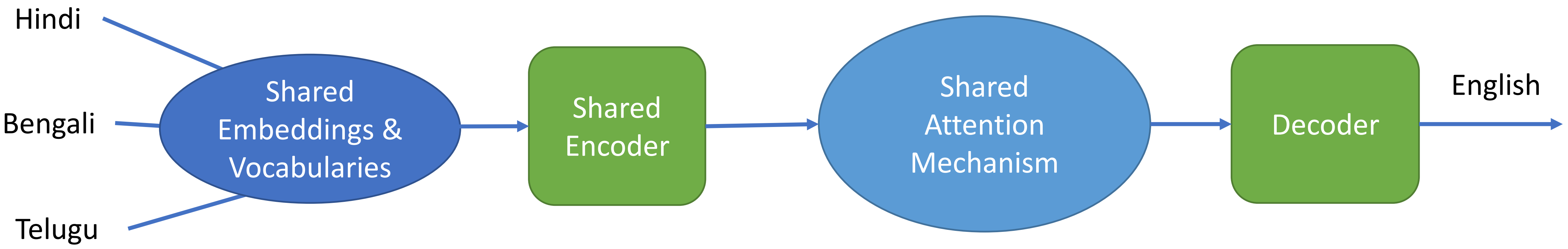
Model tuned for C_1

Zeroshot translation

Can we translate language pairs we have not seen so far?

- Unseen language pair
- Unseen source language
- Unseen target language


Hindi
Bengali
Telugu
Encoder$_1$
Encoder$_2$
Encoder$_3$
Shared Attention Mechanism
Decoder$_1$
Decoder$_2$
English
German



With a shared encoder, unseen source languages can be supported

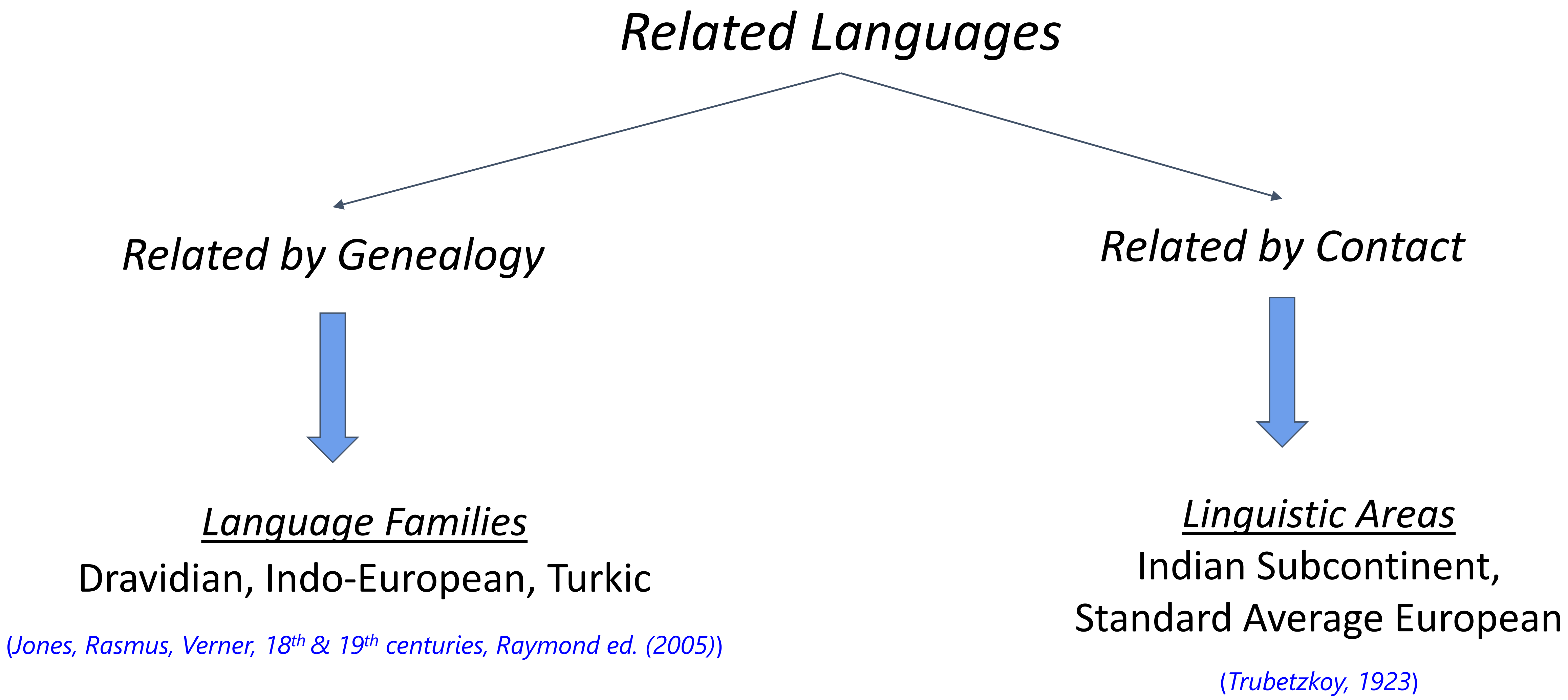
Supporting unseen target languages is a challenge

Outline

- Learning Cross-lingual Embeddings
- Training a Multilingual NLP Application
- Related Languages and Multilingual Learning
- Summary and Research Directions

Related Languages (plus)
Pre-processing Text

Multi-task learning is more beneficial when tasks are related to each other



Related Languages

Related by Genealogy

Related by Contact

<u>Language Families</u>

Dravidian, Indo-European, Turkic

(Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005))

<u>Linguistic Areas</u>

Indian Subcontinent,
Standard Average European

(Trubetzkoy, 1923)

Related languages may not belong to the same language family!

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

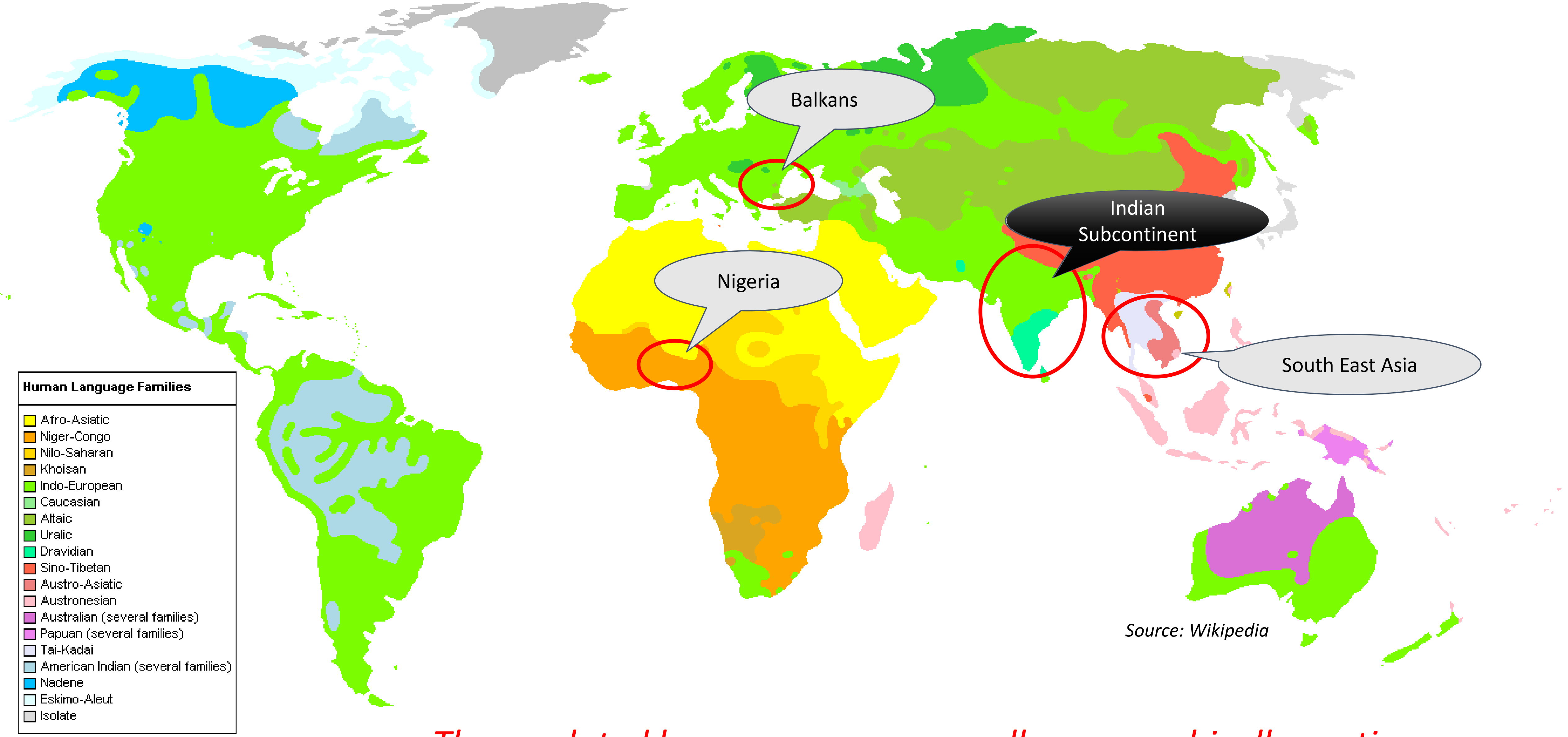
Hindi

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

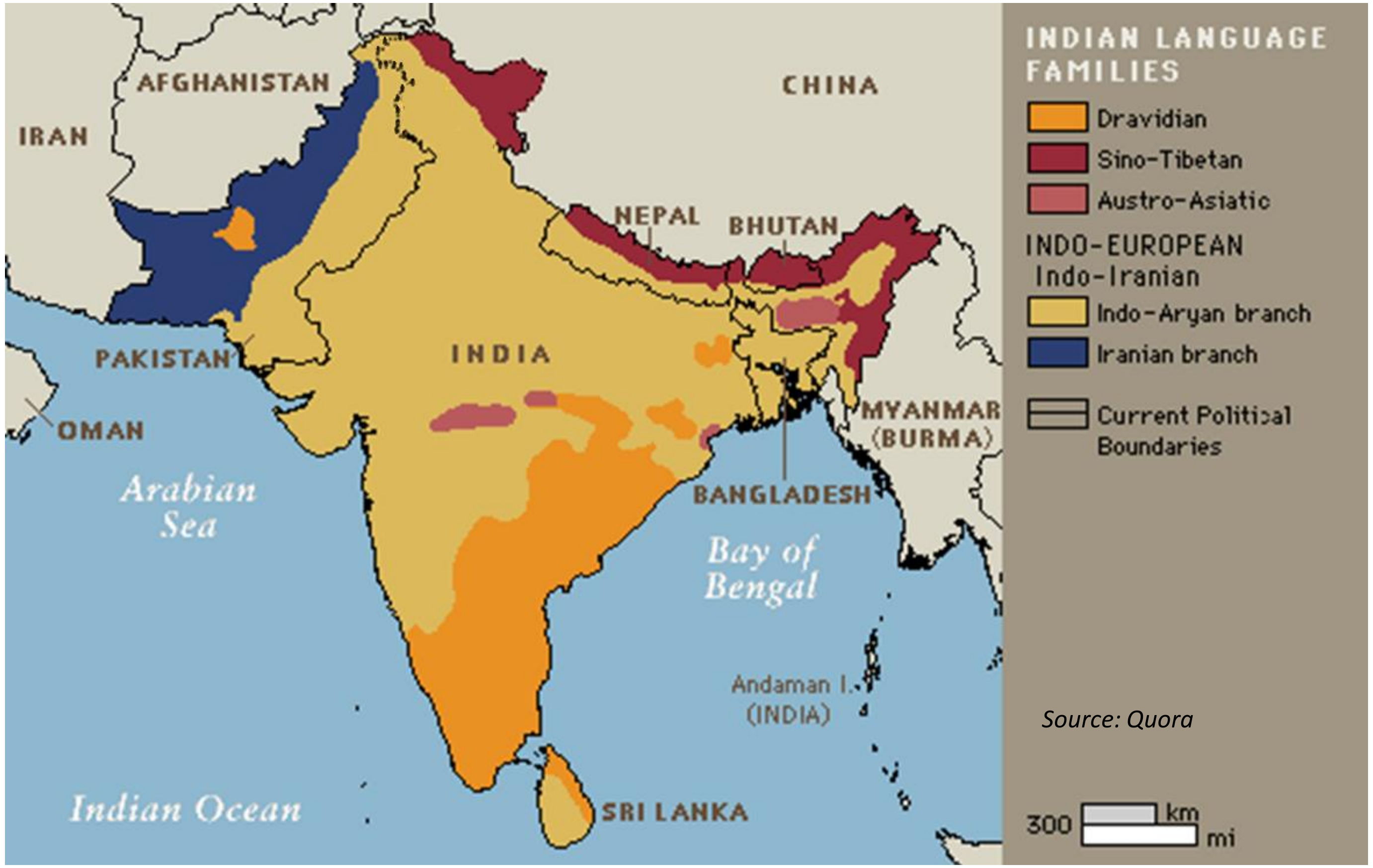
Syntactic: share the same basic word order

Why are we interested in such related languages?



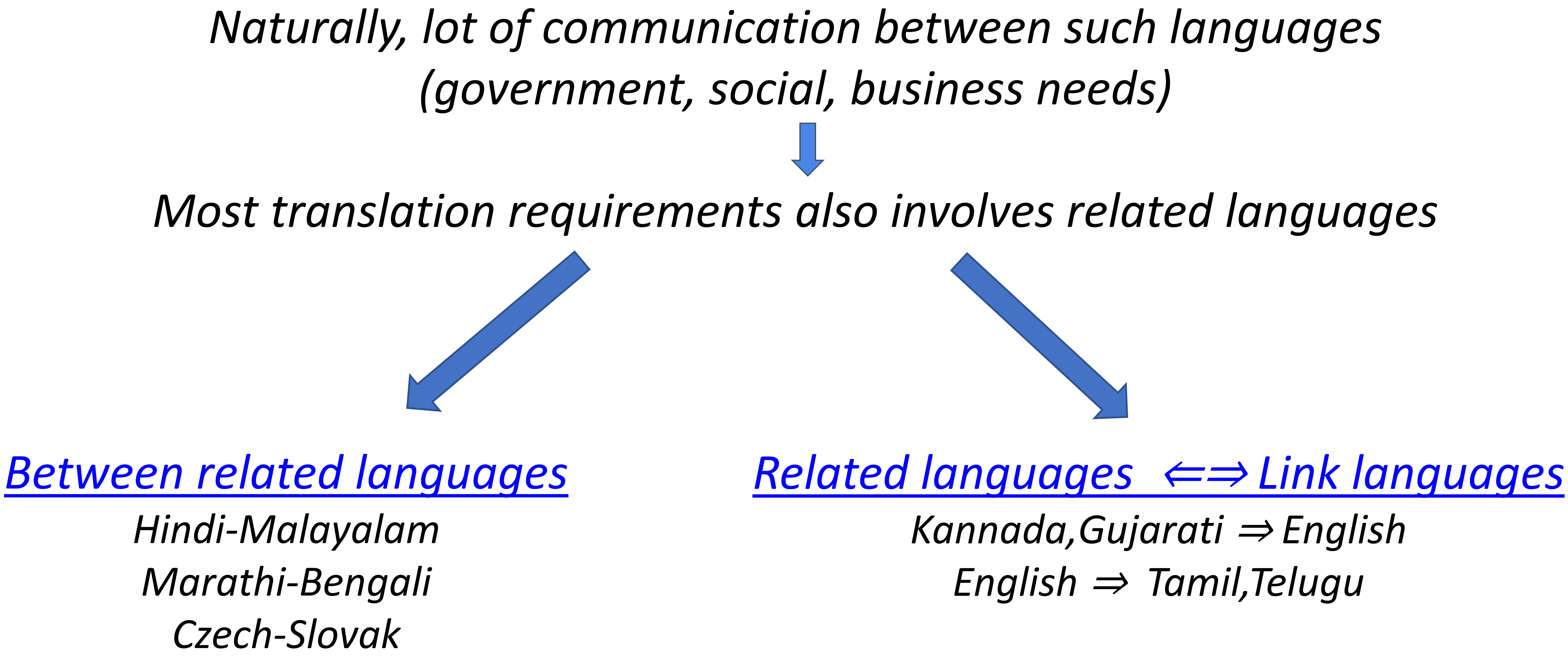
Source: Wikipedia

These related languages are generally geographically contiguous



Source: Quora

- 5 language families (+ 2 to 3 on the Andaman & Nicobar Islands)
- 22 scheduled languages
- 11 languages with more than 25 million speakers
- Highly multilingual country



Naturally, lot of communication between such languages (government, social, business needs)

Most translation requirements also involves related languages

Between related languages

Hindi-Malayalam

Marathi-Bengali

Czech-Slovak

Related languages ⇐⇒ Link languages

Kannada,Gujarati ⇒ English

English ⇒ Tamil,Telugu

We want to be able to handle a large number of such languages

e.g. 30+ languages with a speaker population of 1 million + in the Indian subcontinent

Utilizing Lexical Similarity

Lexically Similar Languages
*(Many words having similar **form** and **meaning**)*

- *Cognates*

a common etymological origin

roTI (hi)	*roTIA (pa)*	*bread*
bhai (hi)	*bhAU (mr)*	*brother*

- *Named Entities*

do not change across languages

mu.mbaI (hi)	mu.mbaI (pa)	mu.mbaI (pa)
keral (hi)	k.eraLA (ml)	keraL (mr)

- *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)	

We want to similar sentences to have similar embeddings

We will find more matches at the sub-word level

Can we use subwords as representation units?

Which subword should we use?

Simple Units of Text Representation

Transliterate unknown words *[Durrani, etal. (2010), Nakov & Tiedemann (2012)]*

(a) Primarily used to handle proper nouns *(b) Limited use of lexical similarity*

स्वातंत्र्य → स्वतंत्रता

Translation of shared lexically similar words can be seen as kind of transliteration

Character *[Vilar, etal. (2007), Tiedemann (2009)]*

Limited context of character level representation

Character n-gram ⇒ increase in data sparsity

Limited benefit ....

... just for closely related languages

Macedonian - Bulgarian, Hindi-Punjabi, etc.

Orthographic Syllable *(Kunchukuttan & Bhattacharyya, 2016a)*

(CONSONANT) + VOWEL

Examples: ca, cae, coo, cra, की (kI), प्रे (pre)

अभिमान ➔ अ भि मा न

Pseudo-Syllable

True Syllable ⇒ Onset, Nucleus and Coda

Orthographic Syllable ⇒ Onset, Nucleus

- Generalization of ***akshara***, the fundamental organizing principle of Indian scripts
- Linguistically motivated, variable length unit
- *Number of syllables in a language is finite*
- Used successfully in transliteration

Byte Pair Encoded (BPE) Unit

(Kunchukuttan & Bhattacharyya, 2017a; Nguyen and Chang, 2017)

- *There may be frequent subsequences in text other than syllables*
- *Herdan-Heap Law ⇒ Syllables are not sufficient*
- *These subsequences may not be valid linguistic units*
- *But they represent statistically important patterns in text*

How do we identify such frequent patterns?

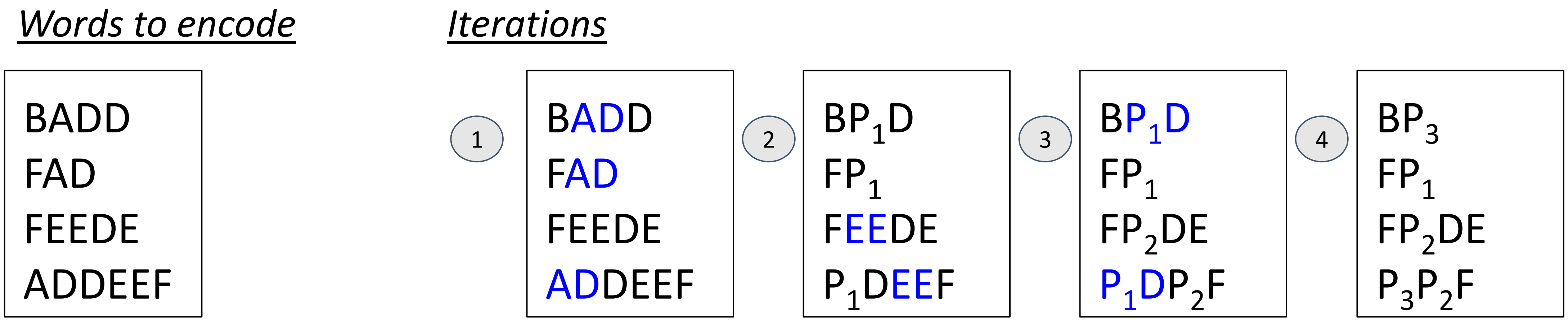
Byte Pair Encoding (Sennrich et al, 2016), Wordpieces (Wu et al, 2016), Huffman encoding based units (Chitnis & DeNero, 2015)

Byte Pair Encoded (BPE) Unit

Byte Pair Encoding is a compression technique (Gage, 1994)

Number of BPE merge operations=3
Vocab: A B C D E F

P_1=AD P_2=EE P_3=P_1D

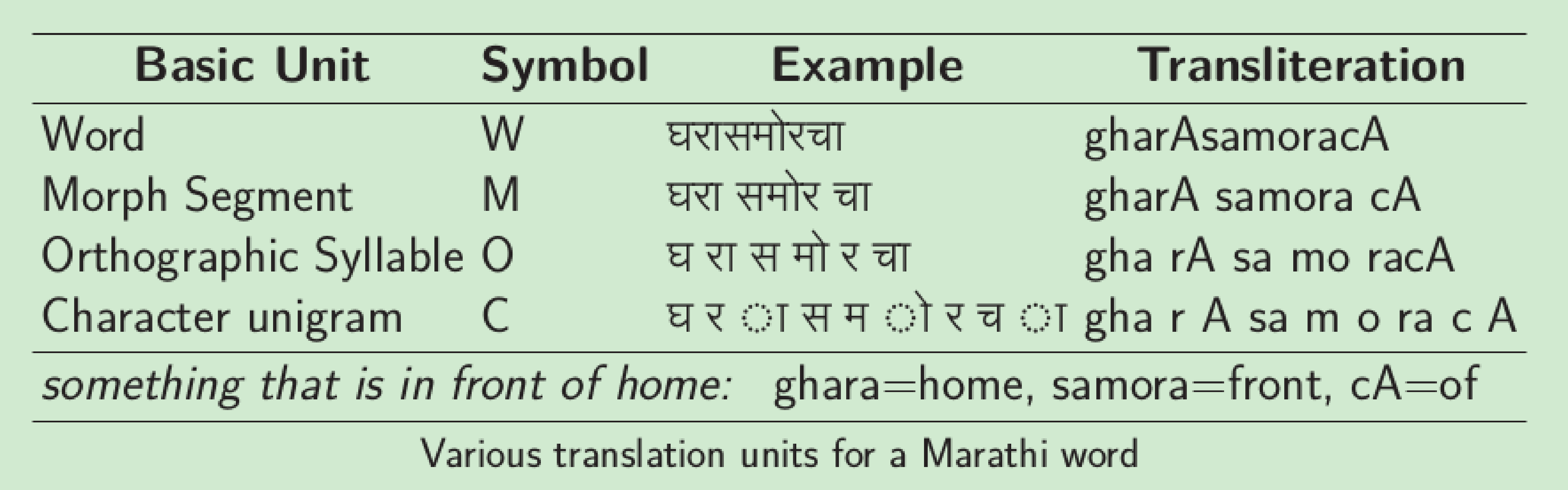


Words to encode	*Iterations*			
	1	2	3	4
BADD	BADD	BP_1D	BP_1D	BP_3
FAD	FAD	FP_1	FP_1	FP_1
FEEDE	FEEDE	FEEDE	FP_2DE	FP_2DE
ADDEEF	ADDEEF	P_1DEEF	P_1DP_2F	P_3P_2F

Data-dependent segmentation

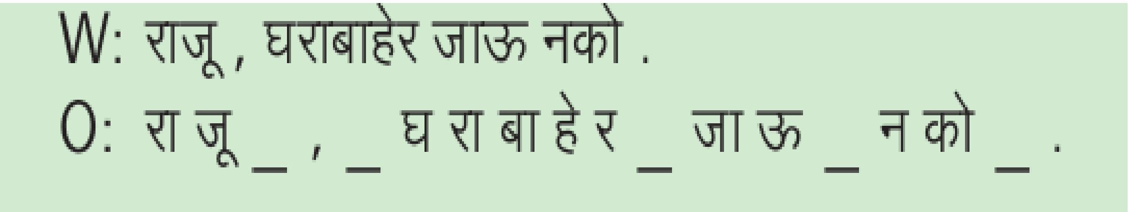
- Inspired from compression theory
- MDL Principle *(Rissansen, 1978)* ⇒ Select segmentation which maximizes data likelihood

Example of various translation units



Basic Unit	Symbol	Example	Transliteration
Word	W	घरासमोरचा	gharAsamoracA
Morph Segment	M	घरा समोर चा	gharA samora cA
Orthographic Syllable	O	घ रा स मो र चा	gha rA sa mo racA
Character unigram	C	घ र ◌ा स म ◌ो र च ◌ा	gha r A sa m o ra c A
something that is in front of home:	ghara=home, samora=front, cA=of		

Various translation units for a Marathi word



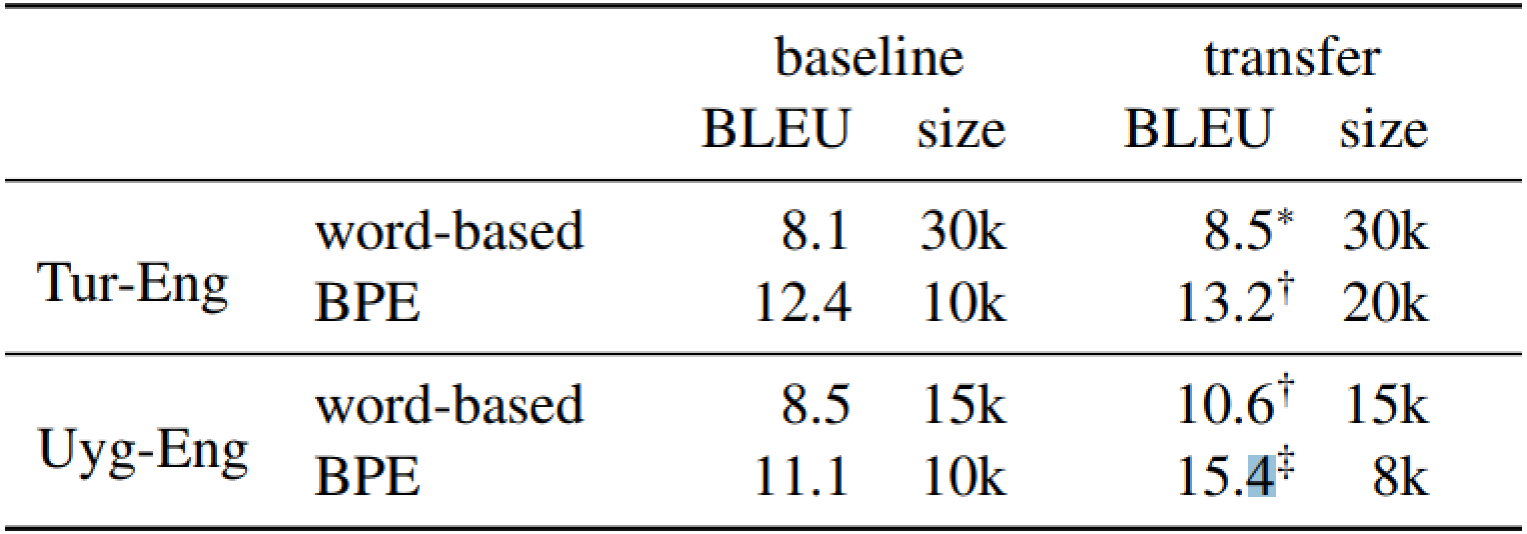
W: राजू , घराबाहेर जाऊ नको .

O: रा जू _ , _ घ रा बा हे र _ जा ऊ _ न को _ .

Instead of a sequence of words, the input to the network is a sequence of subword units

Neural Machine Translation

(Nguyen and Chang, 2017)



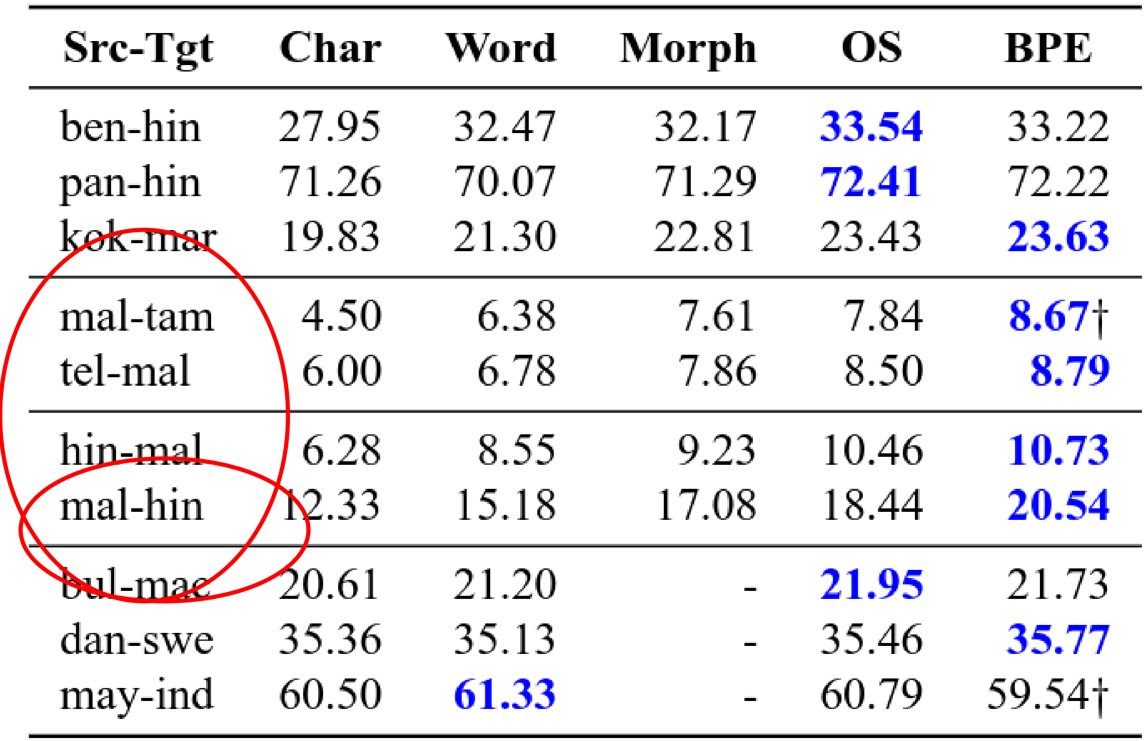
		baseline		transfer	
		BLEU	size	BLEU	size
Tur-Eng	word-based	8.1	30k	8.5^{*}	30k
	BPE	12.4	10k	$13.2^{\dagger}$	20k
Uyg-Eng	word-based	8.5	15k	$10.6^{\dagger}$	15k
	BPE	11.1	10k	$15.4^{\ddagger}$	8k

Uzbek as resource-rich assisting language; Turkish and Uyghur as primary languages

Size: refers to vocabulary size

Statistical Machine Translation

(Kunchukuttan & Bhattacharyya, 2016a; Kunchukuttan & Bhattacharyya, 2017a)

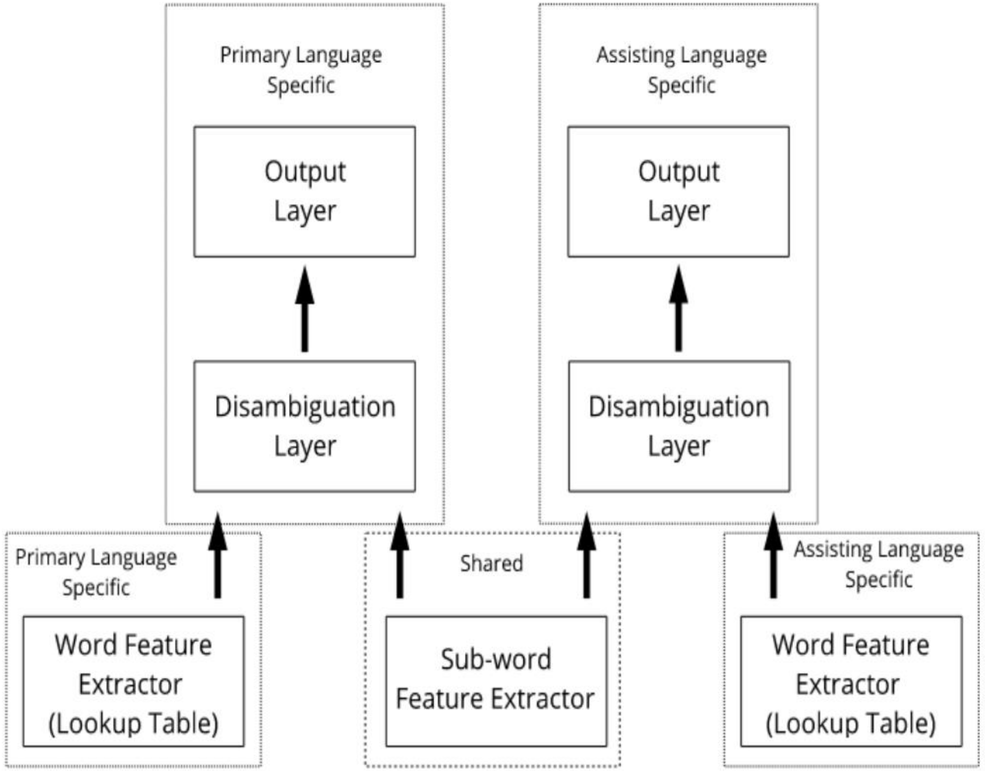


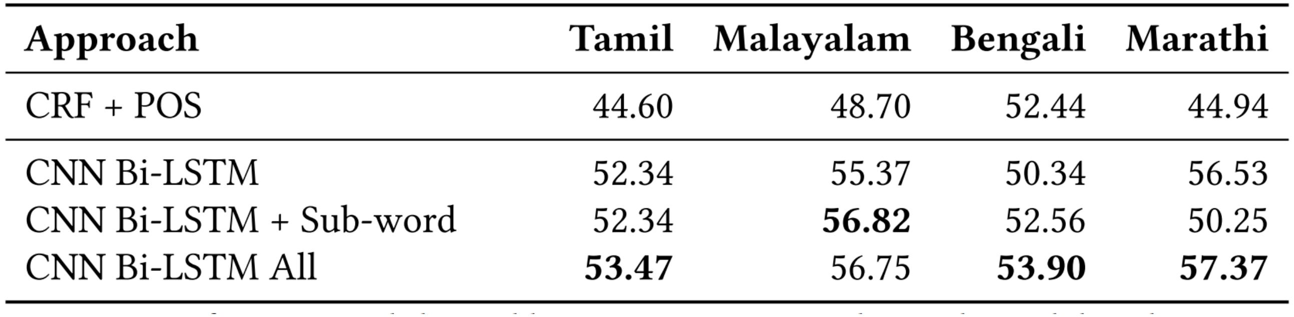
Src-Tgt	Char	Word	Morph	OS	BPE
ben-hin	27.95	32.47	32.17	**33.54**	33.22
pan-hin	71.26	70.07	71.29	**72.41**	72.22
kok-mar	19.83	21.30	22.81	23.43	**23.63**
mal-tam	4.50	6.38	7.61	7.84	**8.67**†
tel-mal	6.00	6.78	7.86	8.50	**8.79**
hin-mal	6.28	8.55	9.23	10.46	**10.73**
mal-hin	12.33	15.18	17.08	18.44	**20.54**
bul-mac	20.61	21.20	-	**21.95**	21.73
dan-swe	35.36	35.13	-	35.46	**35.77**
may-ind	60.50	**61.33**	-	60.79	59.54†

- Substantial improvement over char-level model (27% & 32% for OS and BPE resp.)
- Significant improvement over word and morph level baselines (11-14% and 5-10% resp)
- Improvement even when languages don't belong to same family (contact exists)
- More beneficial when languages are morphologically rich

Named Entity Recognition

(Rudramurthy et al., 2018)


Primary Language Specific
Output Layer
Disambiguation Layer
Assisting Language Specific
Output Layer
Disambiguation Layer
Primary Language Specific
Word Feature Extractor (Lookup Table)
Shared
Sub-word Feature Extractor
Assisting Language Specific
Word Feature Extractor (Lookup Table)



Approach	Tamil	Malayalam	Bengali	Marathi
CRF + POS	44.60	48.70	52.44	44.94
CNN Bi-LSTM	52.34	55.37	50.34	56.53
CNN Bi-LSTM + Sub-word	52.34	**56.82**	52.56	50.25
CNN Bi-LSTM All	**53.47**	56.75	**53.90**	**57.37**

Utilizing Syntactic Similarity

(Kunchukuttan et al., 2014)

Phrase based MT is not good at learning word ordering

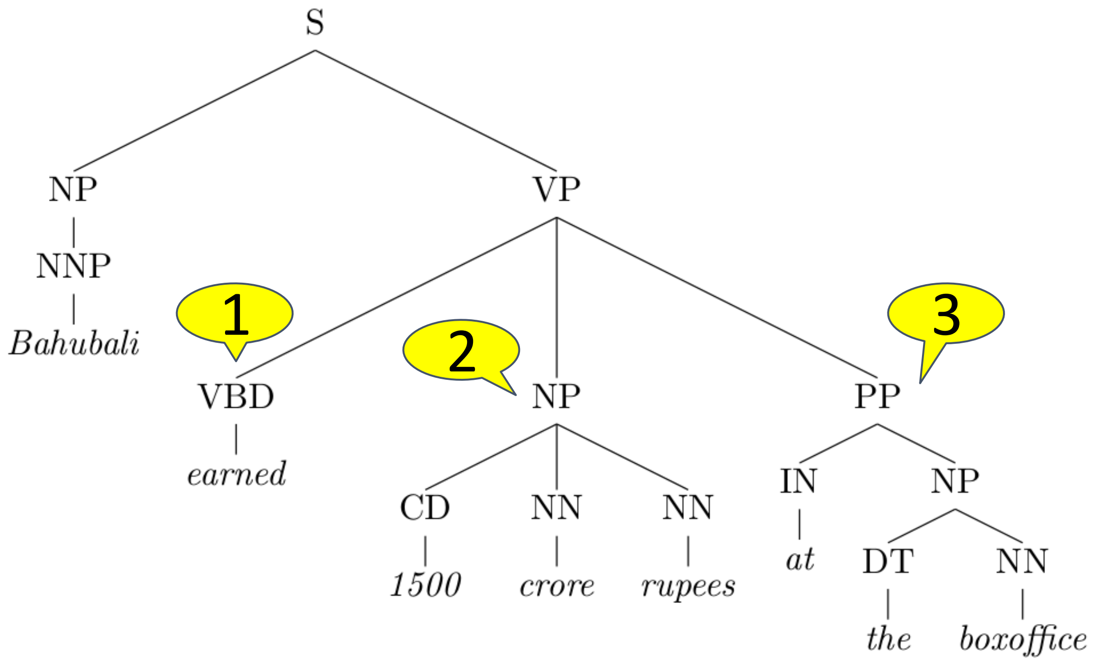
Solution: Let's help PB-SMT with some preprocessing of the input

Change order of words in input sentence to match order of the words in the target language

Let's take an example

Bahubali earned more than 1500 crore rupee sat the boxoffice

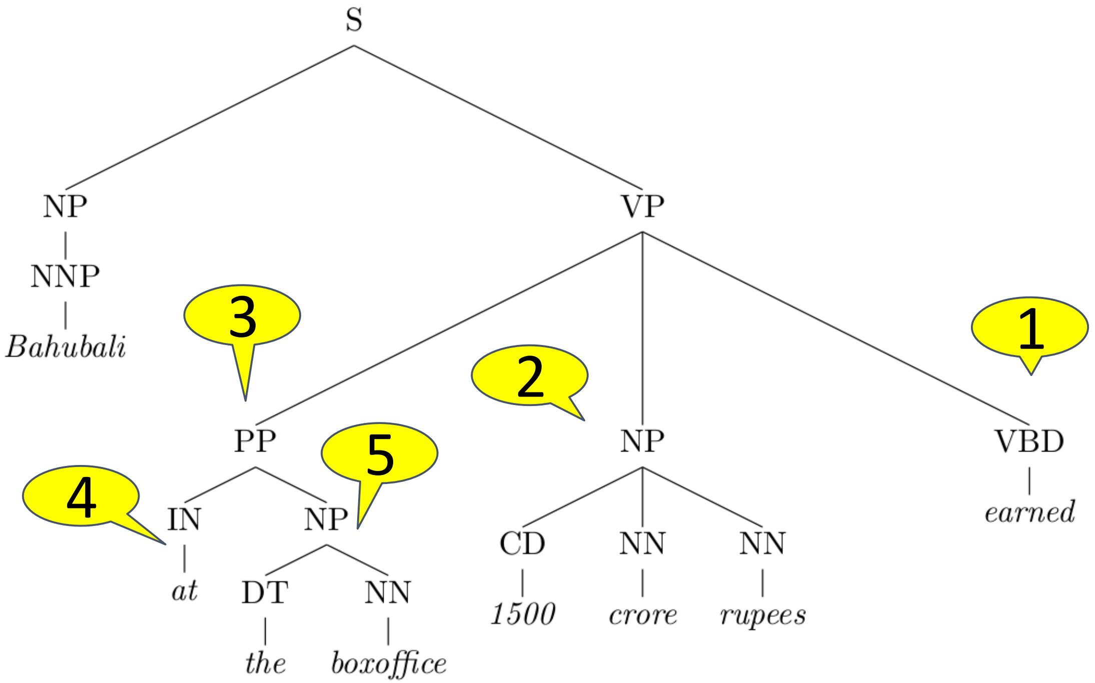
Parse the sentence to understand its syntactic structure



Apply rules to transform the tree

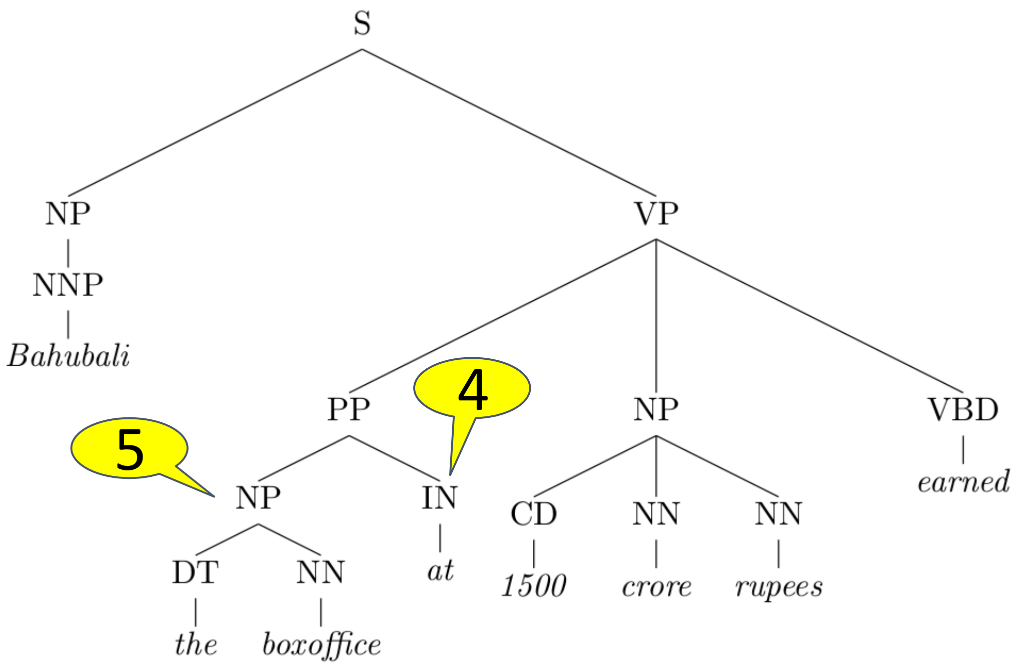
VP → VBD NP PP ⇒ VP → PP NP VBD

This rule captures Subject-Verb-Object to Subject-Object-Verb divergence



Prepositions in English become postpositions in Hindi

PP → IN NP ⇒ PP → NP IN



The new input to the machine translation system is

Bahubali the boxoffice at 1500 crore rupees earned

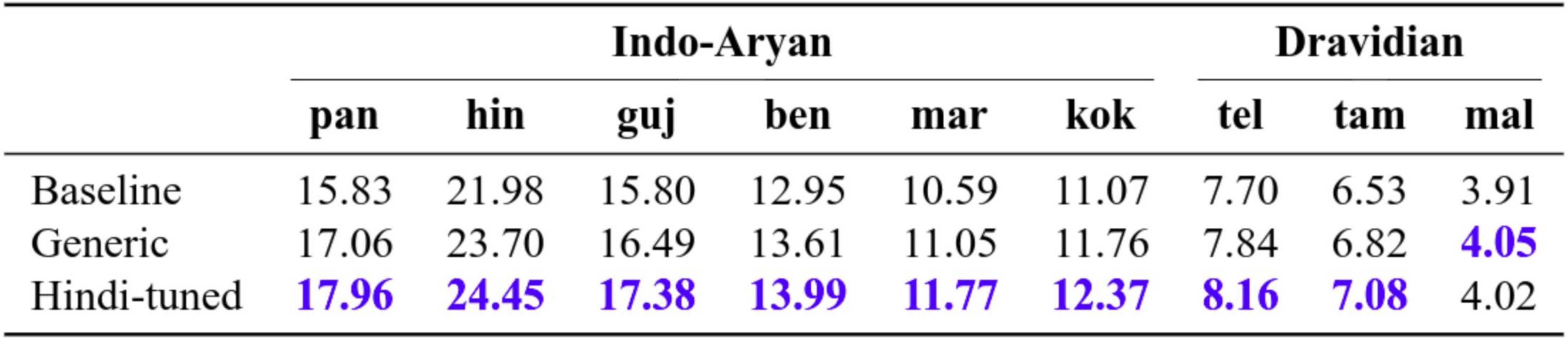
Now we can translate with little reordering

बाहुबली ने बॉक्सओफिस पर 1500 करोड रुपए कमाए

These rules can be written manually or learnt from parse trees

Can we reuse English-Hindi rules for English-Indian languages?

All Indian languages have the same basic word order



	Indo-Aryan						Dravidian		
	pan	**hin**	**guj**	**ben**	**mar**	**kok**	**tel**	**tam**	**mal**
Baseline	15.83	21.98	15.80	12.95	10.59	11.07	7.70	6.53	3.91
Generic	17.06	23.70	16.49	13.61	11.05	11.76	7.84	6.82	**4.05**
Hindi-tuned	**17.96**	**24.45**	**17.38**	**13.99**	**11.77**	**12.37**	**8.16**	**7.08**	4.02

(Kunchukuttan et al., 2014)

Generic reordering *(Ramanathan et al 2008)*

Basic reordering transformation for English→ Indian language translation

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

Improvement over the basic rules by analyzing English → Hindi translation output

Utilizing Orthographic Similarity

Orthographically Similar Languages

(a) *highly overlapping phoneme sets*

(b) *mutually compatible orthographic systems*

(c) *similar grapheme to phoneme mappings*

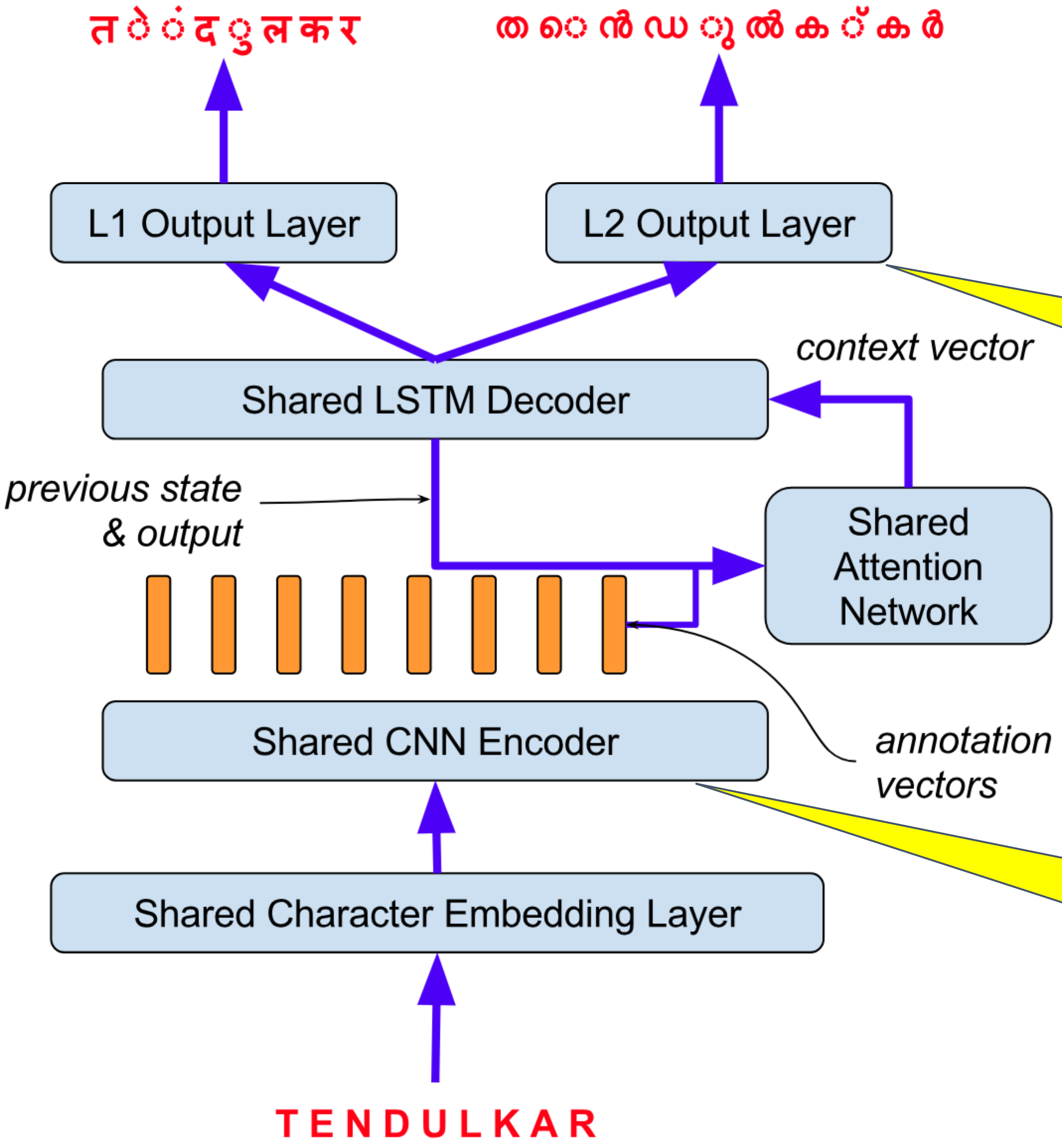
e.g. Indic languages

Can be useful in multilingual settings like:

Transliteration, grapheme to phoneme, Speech recognition , TTS , short text translation for related languages (tweets, headlines),

Multilingual Neural Transliteration

(Kunchukuttan et al., 2018)

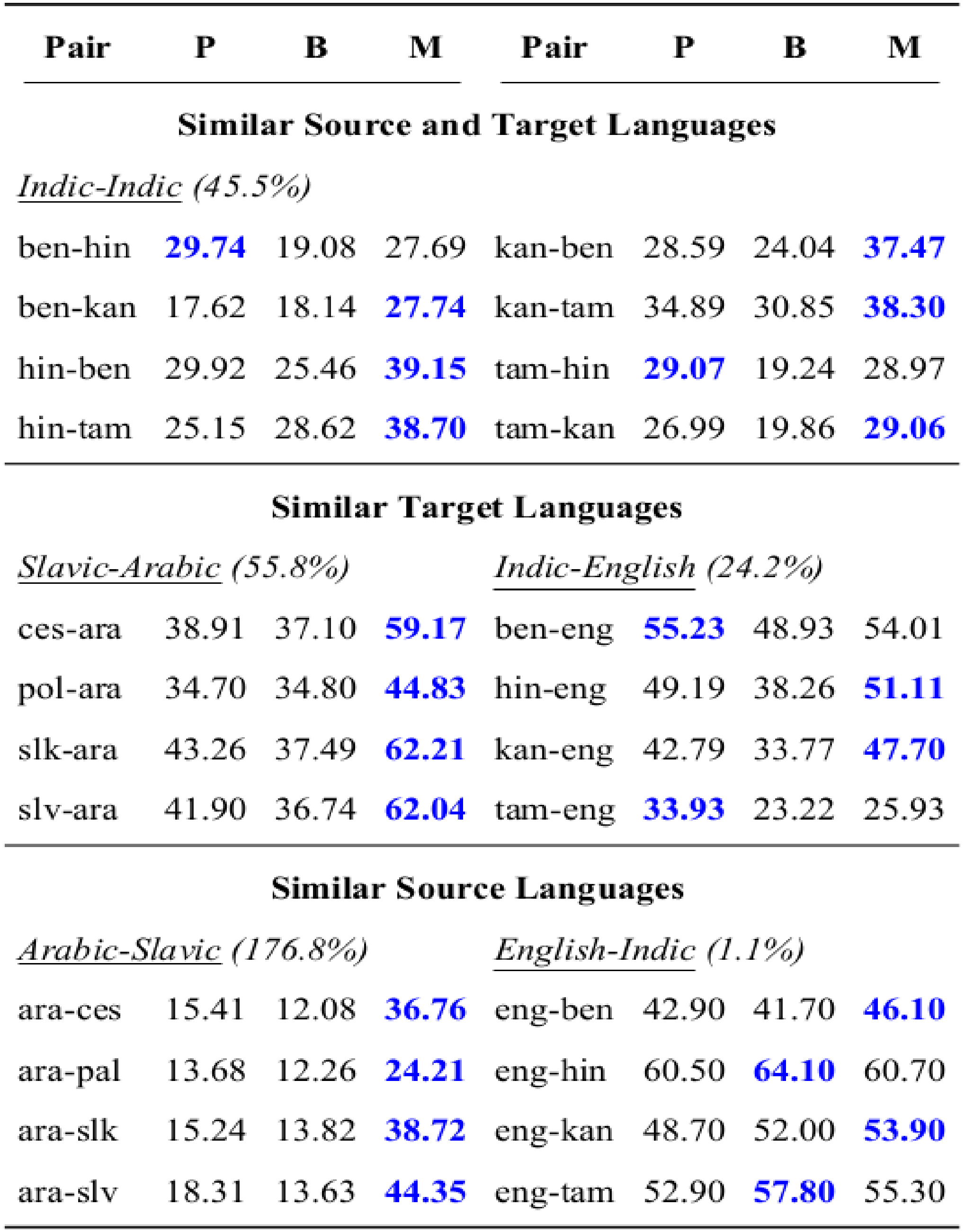


Compact Architecture

Shared embeddings, encoder, decoder and attention layer

Language Specific Output Layer

Compact Network



Pair	P	B	M	Pair	P	B	M
Similar Source and Target Languages							
Indic-Indic (45.5%)							
ben-hin	**29.74**	19.08	27.69	kan-ben	28.59	24.04	**37.47**
ben-kan	17.62	18.14	**27.74**	kan-tam	34.89	30.85	**38.30**
hin-ben	29.92	25.46	**39.15**	tam-hin	**29.07**	19.24	28.97
hin-tam	25.15	28.62	**38.70**	tam-kan	26.99	19.86	**29.06**
Similar Target Languages							
Slavic-Arabic (55.8%)				*Indic-English (24.2%)*			
ces-ara	38.91	37.10	**59.17**	ben-eng	**55.23**	48.93	54.01
pol-ara	34.70	34.80	**44.83**	hin-eng	49.19	38.26	**51.11**
slk-ara	43.26	37.49	**62.21**	kan-eng	42.79	33.77	**47.70**
slv-ara	41.90	36.74	**62.04**	tam-eng	**33.93**	23.22	25.93
Similar Source Languages							
Arabic-Slavic (176.8%)				*English-Indic (1.1%)*			
ara-ces	15.41	12.08	**36.76**	eng-ben	42.90	41.70	**46.10**
ara-pal	13.68	12.26	**24.21**	eng-hin	60.50	**64.10**	60.70
ara-slk	15.24	13.82	**38.72**	eng-kan	48.70	52.00	**53.90**
ara-slv	18.31	13.63	**44.35**	eng-tam	52.90	**57.80**	55.30

Top-1 accuracy for Phrase-based (P), bilingual neural (B) and multilingual neural (P)

Qualitative Analysis

Major reduction in vowel related errors

Reduction in confusion between similar consonants

e.g. (T,D), (P,B)

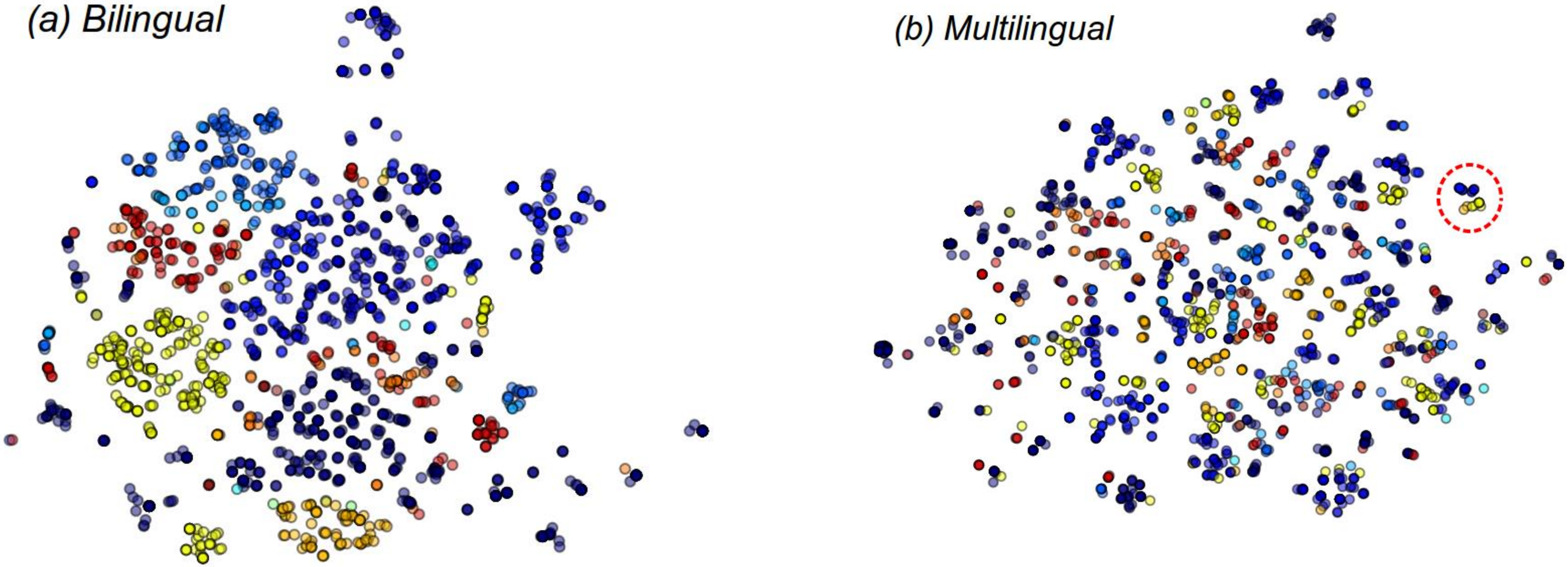
Generates more canonical outputs

For मोरिस, moris is a valid spelling but maurice is canonical

- *May explain less improvement in en-Indic*

Why does Multilingual Training help?

Encoder learns specialized contextual representations



Outline

- Learning Cross-lingual Embeddings
- Training a Multilingual NLP Application
- Related Languages and Multilingual Learning
- Summary and Research Directions

Summary

- Cross-lingual word embeddings are the cornerstone for sharing training data across languages
- Tremendous advances in unsupervised learning of cross-lingual embeddings
- Ensuring word embeddings map to a common space is not sufficient
  - Encoder outputs have to be mapped too
- Related languages can make maximum utilization of task similarity and share data

Research Directions

- Do cross-lingual embeddings work equally well for all languages?
- Cross-lingual contextualized embedding *i.e.* encoder outputs
- Alternative architectures
  - Transformer architecture shown to work better for multilingual NMT
  - Adversarial learning looks promising
- Target side sharing of parameters is under-investigated

Other Reading Material

- Tutorial on *Multilingual Multimodal Language Processing Using Neural Networks.* Mitesh Khapra and Sarath Chandar. NAACL 2016.
- Tutorial on *Cross-Lingual Word Representations: Induction and Evaluation.* Ivan Vuli¢, Anders Søgaard, Manaal Faruqui. EMNLP 2017.
- Tutorial on *Statistical Machine Translation for Related languages. Pushpak Bhattacharyya, Mitesh Khapra, Anoop Kunchukuttan. NAACL 2016.*
- Tutorial on *Statistical Machine Translation and Transliteration for Related languages. Mitesh Khapra, Anoop Kunchukuttan. ICON 2015.*

Tools

- Multilingual Unsupervised and Supervised Embeddings (MUSE)
- VecMap

More pointers in slides from the tutorial Vuli¢, et al., (2017) Slides: https://www.cse.iitb.ac.in/~anoopk/publications/presentations/iiit-ml-multilingual-2018.pdf

Thank you!

Multilingual data, code for Indian languages

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

https://www.cse.iitb.ac.in/~anoopk

Work with Prof. Pushpak Bhattacharyya, Prof. Mitesh Khapra, Abhijit Mishra, Ratish Puduppully, Rajen Chatterjee, Ritesh Shah, Maulik Shah, Pradyot Prakash, Gurneet Singh, Raj Dabre, Rohit More, Rudramurthy

References

1. Abbi, A. (2012). Languages of india and india and as a linguistic area. http://www.andamanese.net/LanguagesofIndiaandIndiaasalinguisticarea.pdf. Retrieved November 15, 2015.
2. Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. In ACL.
3. Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2289--2294, Austin, Texas. Association for Computational Linguistics.
4. Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 451--462. Association for Computational Linguistics.
5. Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pages 5012--5019.
6. Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
7. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. ICLR 2015.
8. Caruana, R. (1997). Multitask learning. Machine learning.
9. Chandar, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In Advances in Neural Information Processing Systems, pages 1853--1861.
10. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In International Conference on Learning Representations.
11. De Saussure, F. (1916). Course in general linguistics. Columbia University Press.
12. Dinu, G., Lazaridou, A., and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In ICLR.
13. Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In Annual Meeting of the Association for Computational Linguistics.
14. Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2017). Multilingual training of crosslingual word embeddings. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers, pages 894--904.

References

15. Durrani, N., Sajjad, H., Fraser, A., and Schmid, H. (2010). Hindi-to-urdu machine translation through transliteration. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.
16. Emeneau, M. B. (1956). India as a Lingustic area. Language.
17. Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 462--471.
18. Finch, A., Liu, L., Wang, X., and Sumita, E. (2015). Neural network transduction models in transliteration generation. In Proceedings of the Fifth Named Entities Workshop (NEWS).
19. Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In Conference of the North American Chapter of the Association for Computational Linguistics.
20. Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. (2016). Multilingual language processing from bytes. NAACL.
21. Gispert, A. D. and Marino, J. B. (2006). Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In In Proc. of 5th International Conference on Language Resources and Evaluation (LREC).
22. Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In International Conference on Machine Learning, pages 748--756.
23. Gordon, R. G., Grimes, B. F., et al. (2005). Ethnologue: Languages of the world, volume 15. SIL International Dallas, TX.
24. Grave, E., Joulin, A., and Berthet, Q. (2018). Unsupervised alignment of embeddings with wasserstein procrustes. CoRR, abs/1805.11222.
25. Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. NAACL.
26. Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In ACL.
27. Hoshen, Y. and Wolf, L. (2018). An iterative closest point method for unsupervised word translation. CoRR, abs/1801.06126.
28. Huang, K., Gardner, M., Papalexakis, E. E., Faloutsos, C., Sidiropoulos, N. D., Mitchell, T. M., Talukdar, P. P., and Fu, X. (2015). Translation invariant word embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pages 1084--1088.

References

29. Jha, G. N. (2012). The TDIL program and the Indian Language Corpora Initiative. In Language Resources and Evaluation Conference.
30. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. arXiv preprint arXiv:1611.04558.
31. Joulin, A., Bojanowski, P., Mikolov, T., and Grave, E. (2018). Improving supervised bilingual mapping of word embeddings. CoRR, abs/1804.07745.
32. Joty, S., Nakov, P., Màrquez, L., & Jaradat, I. (2017). Cross-language Learning with Adversarial Neural Networks: Application to Community Question Answering. CoNLL.
33. Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In Proceedings of COLING 2012, pages 1459--1474.
34. Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R., & Bhattacharyya, P. (2014). Sata-anuvadak: Tackling multiway translation of indian languages. LREC.
35. Kunchukuttan, A., & Bhattacharyya, P. (2016a). Orthographic syllable as basic unit for smt between related languages. EMNLP.
36. Kunchukuttan, A., & Bhattacharyya, P. (2016b). Faster decoding for subword level Phrase-based SMT between related languages. VarDIAL.
37. Kunchukuttan, A., & Bhattacharyya, P. (2017a). Learning variable length units for SMT between related languages via Byte Pair Encoding. SCLeM.
38. Kunchukuttan, A., Shah, M., Prakash, P., & Bhattacharyya, P. (2017b). Utilizing Lexical Similarity between Related, Low-resource Languages for Pivot-based SMT. IJCNLP.
39. Kunchukuttan, A., Khapra, M., Singh, G., & Bhattacharyya, P. (2018). Leveraging Orthographic Similarity for Multilingual Neural Transliteration. Transactions Of The Association For Computational Linguistics, 6, 303-316.
40. Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In International Conference on Learning Representations.
41. Lazaridou, A., Dinu, G., and Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 270--280.

References

42. Lee, J., Cho, K., and Hofmann, T. (2017). Fully Character-Level Neural Machine Translation without Explicit Segmentation. Transactions of the Association for Computational Linguistics.
43. Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
44. Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.
45. Nguyen, T. Q., & Chiang, D. (2017). Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. IJCNLP.
46. Rudramurthy, V., Kunchukuttan, A., & Bhattacharyya, P. (2018). Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 401-406).
47. Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. Psychometrika, 31(1):1--10.
48. Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In ACL.
49. Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In ICLR.
50. Subbārāo, K. V. (2012). South Asian languages: A syntactic typology. Cambridge University Press.
51. Tiedemann, J. (2009a). Character-based PBSMT for closely related languages. In Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009).
52. Tiedemann, J. (2009b). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In Recent Advances in Natural Language Processing.
53. Tiedemann, J. and Nakov, P. (2013). Analyzing the use of character-level translation with sparse and noisy datasets. In Recent Advances in Natural Language Processing.
54. Trubetzkoy, N. (1928). Proposition 16. In Actes du premier congres international des linguistes à La Haye.
55. Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In HLT-NAACL, pages 484–491.
56. Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In Proceedings of the Second Workshop on Statistical Machine Translation.
57. Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM.

References

58. Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. Machine Translation, 21(3):165–181.
59. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., and Norouzi, M. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. ArXiv e-prints: abs/1609.08144.
60. Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pages 1006--1011.
61. Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-task cross-lingual sequence tagging from scratch. arXiv preprint arXiv:1603.06270.
62. Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017a). Adversarial training for unsupervised bilingual lexicon induction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1959--1970. Association for Computational Linguistics.
63. Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017b). Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In EMNLP.
64. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. EMNLP.