Unsupervised Neural Machine Translation

Tutorial @ ICON-2020
IIT Patna

Diptesh Kanojia

Jyotsana Khatri

Tamali Banerjee

Prof. Pushpak Bhattacharyya

Rudra Murthy





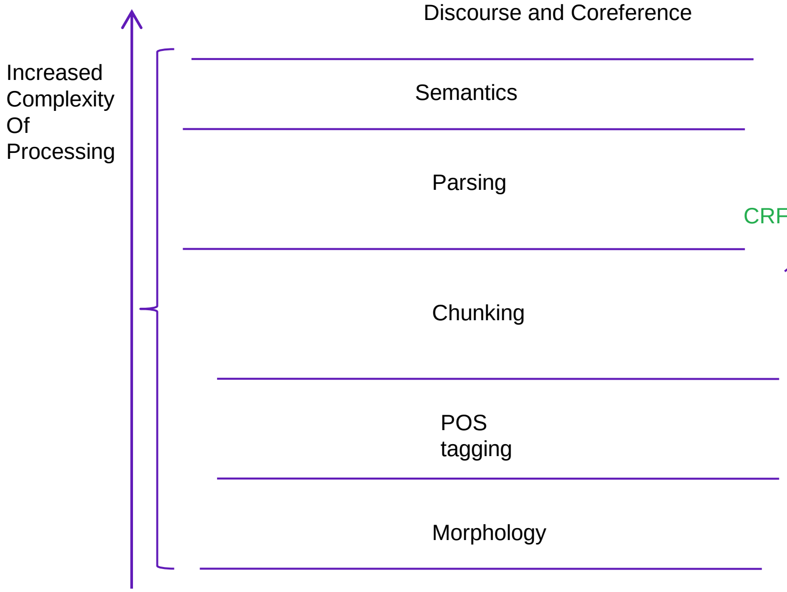
Paradigms of Machine Translation

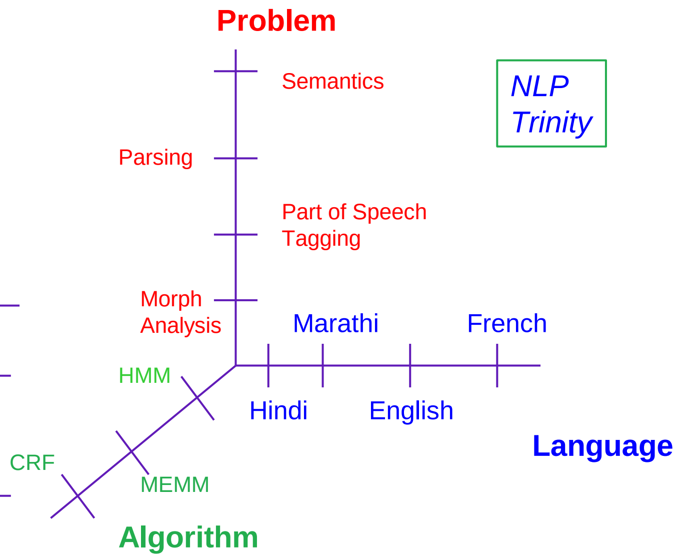
Pushpak Bhattacharyya

Acknowledgement: Numerous PhD, masters and UG students and research staff working on MT with me since 2000

Perspective

NLP: a useful view



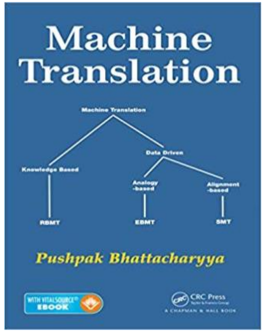


Machine Translation: **Translating** from one language to another by **computer**

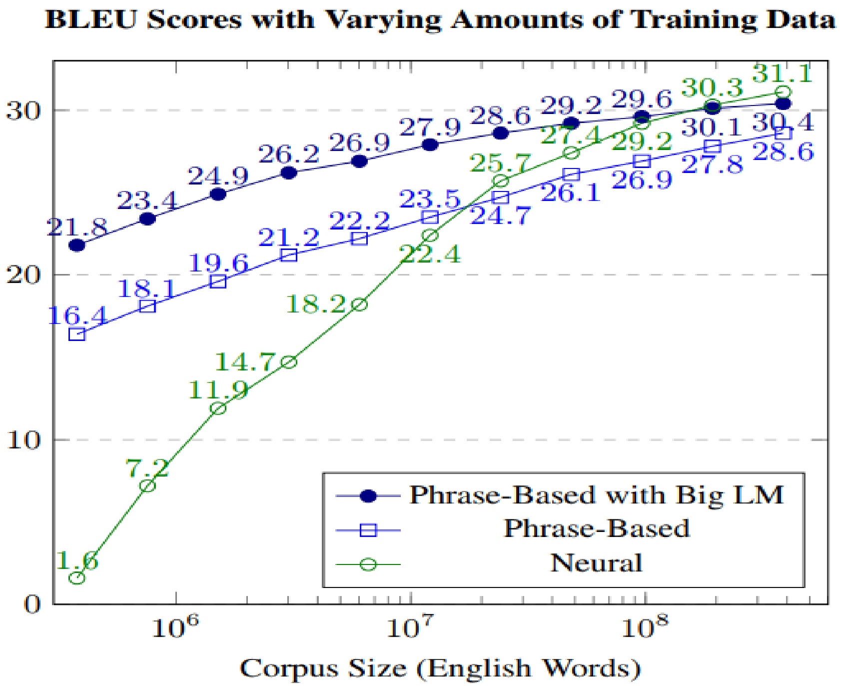
Machine Translation

- Knowledge Based: Essence- **A**nalysis
  - RBMT (rule based)
    - Interlingua based
    - transfer based
- Data Driven
  - Similarity based: Essence- **A**nalogy
    - EBMT (example based)
  - Probability based: Essence- **A**lignment
    - SMT (statistical)
  - Distributed Representation based: Essence- **A**ttention?
    - NMT (neural)

<u>RBMT</u>- rule based MT
<u>EBMT</u>- example based MT
<u>SMT</u>- Statistical MT
<u>NMT</u>- Neural MT



Today's Ruling Paradigm: NMT which is data intensive



Philipp Koehn and Rebecca Knowles. 2017. *Six Challenges for Neural Machine Translation.* In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

Essential Elements of MT Paradigms

- **A**nalysis in RBMT
- **A**lignment in SMT
- **A**nalogy in EBMT
- **A**ttention in NMT?

Challenge of MT: Language Divergence

- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

Our work on English-IL Language Divergence with illustrations from Hindi
(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)

Different ways of expressing meaning

English:

This blanket is very soft

Hindi:

yaha kambal bahut naram hai

Bangla:

ei kambal ti khub naram <null>

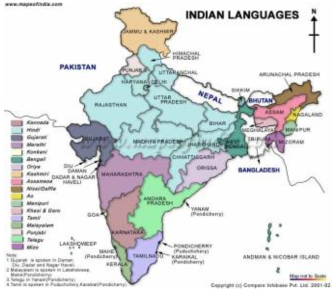
Marathi:

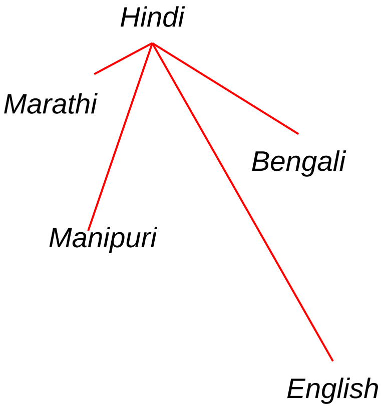
haa kambal khup naram aahe

Manipuri:

kampor asi mon mon laui

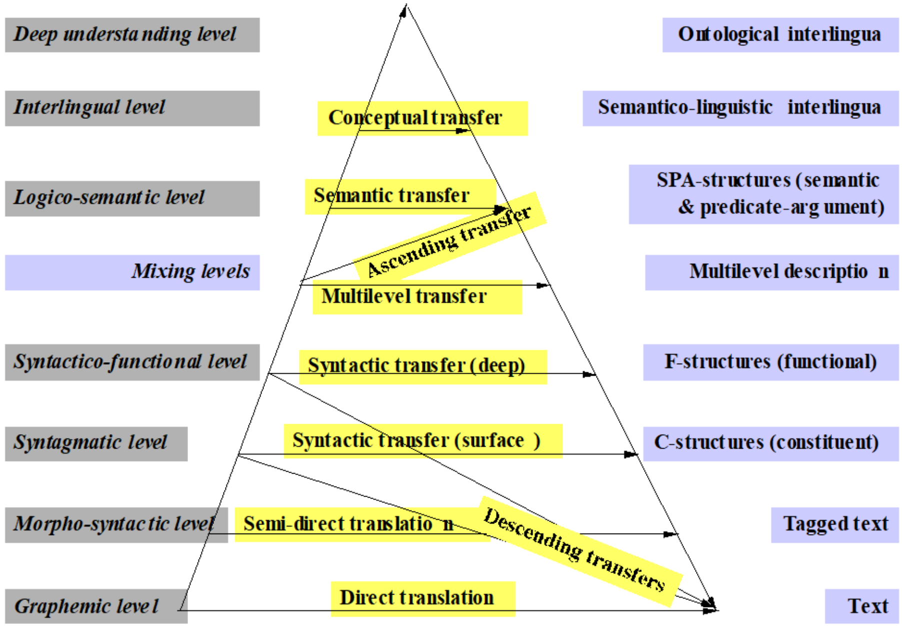
blanket this soft soft is





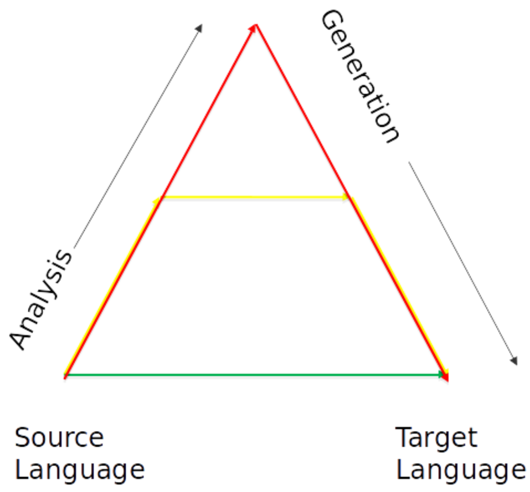
Kinds of MT Systems

(point of entry from source to the target text)



(Vauquois. 1968)

Simplified Vauquois



Interlingua Based Translation

Transfer Based Translation

Direct Translation

Differentiating Interlingual and Transfer based MT: ***TBMT can choose the level of transfer!*** Need to emphasise this point

- राजा को नमन करो (Hindi; Indo Aryan)
- raajaa ko naman karo
- HG: king to obeisance do
- **Give obeisance to the king** (English; Indo-Aryan)

- राजाला नमन करा (Marathi; Indo Aryan)
- raajaalaa naman karaa
- king_to obeisance do

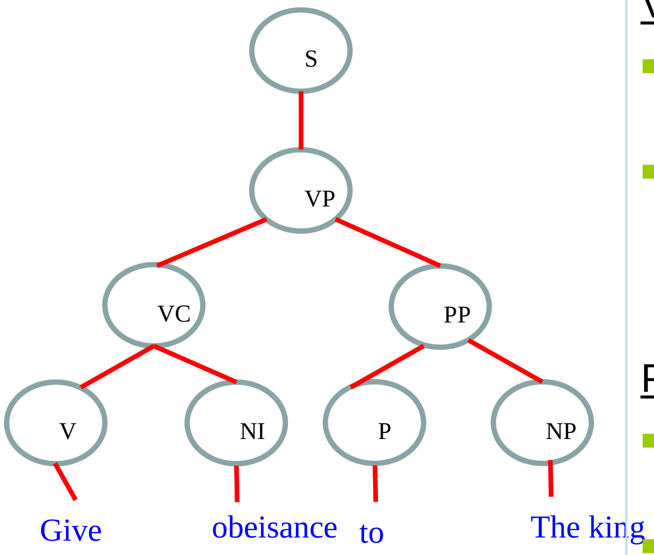
- அரசரை வணங்கு (Tamil; Dravidian)
- aracarai vanaNku
- king_to obeisance_do

- নিংথৌবু খইরম্মু (Manipuri; Tibeto Burman)
- niNgthoubu khoirammu
- king_to obeisance do

transfer amongst different language families

Language	Inflected Verb/Inflected verb complex	Inflected Noun/Inflected Noun chunk
English	*give obeisance*	*To the king*
Hindi	*naman karo*	*raajaa ko*
Marathi	*naman karaa*	*raajaalaa*
Tamil	*vanaNku*	*aracarai*
Manipuri	*Khoirammu*	*niNgthoubu*

English parse tree



Transfer rules:

- VC-PP inversion (all languages)

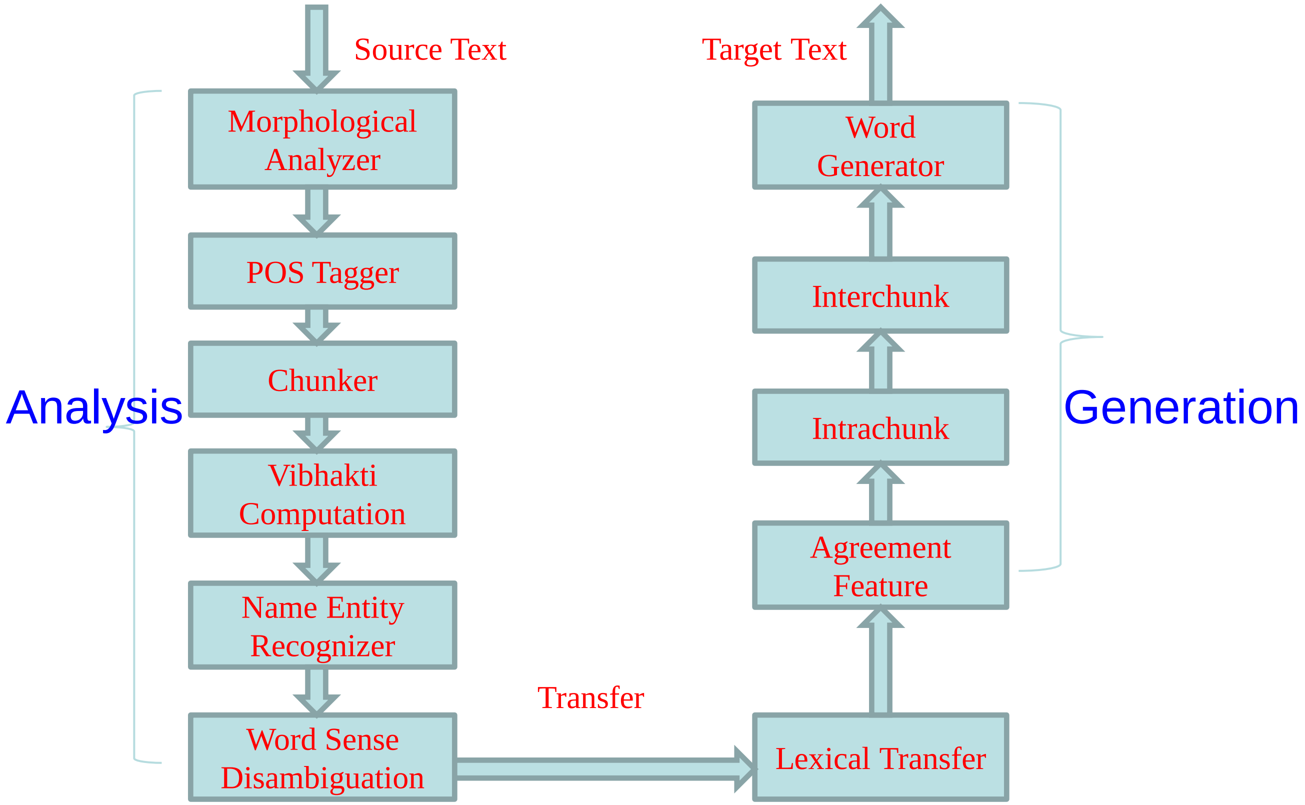
<u>VC</u>

- V-NI inversion (H & M: naman karo, naman karaa)
- V-NI combination → nominal verb with appropriate inflection (T, Mn: vanaNku, khoirammu)

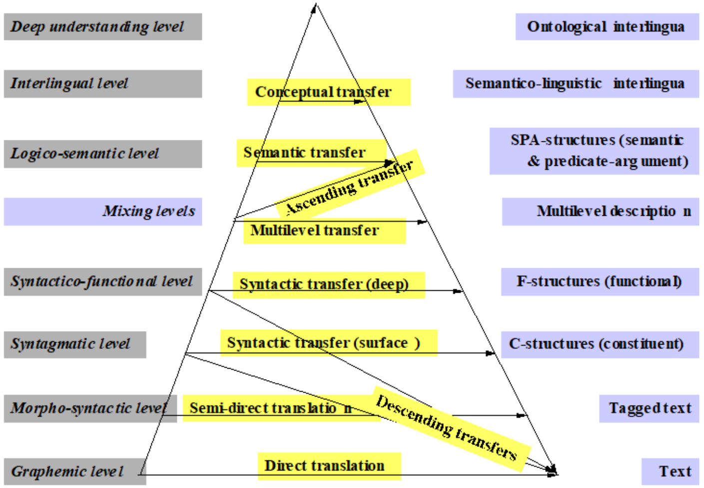
<u>PP</u>

- PP inversion with P becoming a postposition (H: raajaa ko)
- suffixed form of ‘king’ expressing accusative case (M, T, Mn: raajaalaa, aracarai, niNgthoubu)

Rule based MT (typical architecture)



Statistical Machine Translation



Foundation

- Data driven approach
- Goal is to find out the English sentence *e* given foreign language sentence *f* whose *p(e|f)* is maximum.

$$\tilde{e} = \underset{e \in e^*}{\operatorname{argmax}} \, p(e|f) = \underset{e \in e^*}{\operatorname{argmax}} \; p(f|e)p(e)$$

- Translations are generated on the basis of statistical model
- Parameters are estimated using bilingual parallel corpora

SMT: Language Model

- To detect *good* English sentences
- Probability of an English sentence $w_1 w_2 \ldots\ldots w_n$ can be written as

  $$Pr(w_1 w_2 \ldots\ldots w_n) = Pr(w_1) * Pr(w_2|w_1) * \ldots * Pr(w_n|w_1\, w_2 \ldots w_{n-1})$$

- Here $Pr(w_n|w_1\, w_2 \ldots w_{n-1})$ is the probability that word w_n follows word string $w_1\, w_2 \ldots w_{n-1}$.
  - N-gram model probability
- Trigram model probability calculation

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

SMT: Translation Model

- *P*(*f*|*e*): Probability of some *f* given hypothesis English translation *e*
- How to assign the values to *p*(*e*|*f*) ?

$$p(f|e) = \frac{count(f,e)}{count(e)} \longleftarrow \text{Sentence level}$$

  - Sentences are infinite, not possible to find pair(e,f) for all sentences

- Introduce a hidden variable ***a***, that represents alignments between the individual words in the sentence pair

$$\Pr(\boldsymbol{f}|\boldsymbol{e}) = \sum_{\boldsymbol{a}} \Pr(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) \longleftarrow \text{Word level}$$

Alignment

- If the string, $e= e_1^l= e_1\ e_2\ \ldots e_l$, has l words, and the string, $f= f_1^m=f_1 f_2 \ldots f_m$, has m words,
- then the alignment, a, can be represented by a series, $\boldsymbol{a_1^m= a_1 a_2 \ldots a_m}$, of m values, each between 0 and l such that if the word in position j of the f-string is connected to the word in position i of the e-string, then
  - $\boldsymbol{a_j= i}$, and
  - if it is not connected to any English word, then $\boldsymbol{a_j= O}$

Example of alignment

English: *Ram went to school*

Hindi: *raam paathashaalaa gayaa*

Ram *went* *to* *school*

<Null> *raam* *paathashaalaa* *gayaa*

Translation Model: Exact expression

$$\Pr(\boldsymbol{f}, \boldsymbol{a} | \boldsymbol{e}) = \Pr(m|\boldsymbol{e}) \prod_{j=1}^{m} \Pr\left(a_j \,|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}\right) \Pr\left(f_j | a_1^{j}, f_1^{j-1}, m, \boldsymbol{e}\right)$$

Choose the length of foreign language string given *e*

Choose alignment given *e* and *m*

Choose the identity of foreign word given *e, m, a*

- Five models for estimating parameters in the expression [2]
- Model-1, Model-2, Model-3, Model-4, Model-5

Proof of Translation Model: Exact expression

$$\Pr(f \mid e) = \sum_a \Pr(f, a \mid e)$$; marginalization

$$\Pr(f, a \mid e) = \sum_m \Pr(f, a, m \mid e)$$; marginalization

$$\Pr(f, a, m \mid e) = \sum_m \Pr(m \mid e) \Pr(f, a \mid m, e)$$

$$= \sum_m \Pr(m \mid e) \Pr(f, a \mid m, e)$$

$$= \sum_m \Pr(m \mid e) \prod_{j=1}^{m} \Pr(f_j, a_j \mid a_1^{j-1}, f_1^{j-1}, m, e)$$

$$= \sum_m \Pr(m \mid e) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, e) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, e)$$

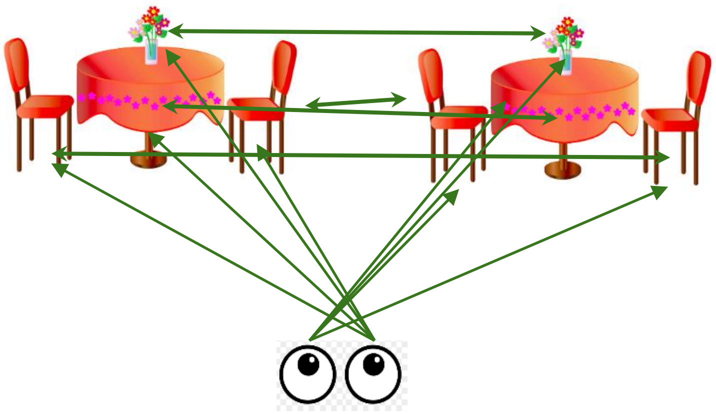
m is fixed for a particular *f*, hence

$$\Pr(f, a, m \mid e) = \Pr(m \mid e) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, e) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, e)$$

Alignment

How to build part alignment from whole alignment

- Two images are in alignment: images on the two retina
- Need to find alignment of parts of it



Fundamental and ubiquitous

- Spell checking
- Translation
- Transliteration
- Speech to text
- Text to Speech

The all important **word alignment**

- The edifice on which the structure of SMT is built (Brown et. Al., 1990, 1993; Och and Ney, 1993)
- Word alignment → Phrase alignment (Koehn et al, 2003)
- Word alignment → Tree Alignment (Chiang 2005, 2008; Koehn 2010)
- Alignment at the heart of Factor based SMT too (Koehn and Hoang 2007)

EM for word alignment from sentence alignment: example

English	French
(1) three rabbits	(1) trois lapins
a b	w x
(2) rabbits of Grenoble	(2) lapins de Grenoble
b c d	x y z

Initial Probabilities:
each cell denotes *t(a←→w), t(a←→x) etc.*

	a	b	c	d
w	1/4	1/4	1/4	1/4
x	1/4	1/4	1/4	1/4
y	1/4	1/4	1/4	1/4
z	1/4	1/4	1/4	1/4

Example of expected count

C[w←→a; (a b)←→(w x)]

$$= \frac{t(w \leftarrow\rightarrow a)}{t(w \leftarrow\rightarrow a)+t(w \leftarrow\rightarrow b)} \text{ X } \#(a \text{ in } `a\ b') \text{ X } \#(w \text{ in } `w\ x')$$

$$= \frac{1/4}{1/4+1/4} \text{ X } 1 \text{ X } 1 = 1/2$$

“counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	1/2	0	0
x	1/2	1/2	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	1/3	1/3	1/3
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Revised probability: example

$$t_{revised}(a \leftarrow\rightarrow w)$$

$$= \frac{1/2}{(1/2+1/2+0+0)_{(a\ b)\leftarrow\rightarrow(w\ x)} + (0+0+0+0)_{(b\ c\ d)\leftarrow\rightarrow(x\ y\ z)}}$$

Revised probabilities table

	a	b	c	d
w	1/2	1/2	0	0
x	1/4	5/12	1/6	1/6
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

“revised counts”

a b ↔ ***w x***	a	b	c	d
w	1/2	3/8	0	0
x	1/2	5/8	0	0
y	0	0	0	0
z	0	0	0	0

b c d ↔ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	5/9	1/3	1/3
y	0	2/9	1/3	1/3
z	0	2/9	1/3	1/3

Re-Revised probabilities table

	a	b	c	d
w	1/2	1/2	0	0
x	3/16	**85/144**	1/9	1/9
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Continue until convergence; notice that (b,x) binding gets progressively stronger; b=rabbits, x=lapins

Derivation of EM based Alignment Expressions

$$V_E = \text{vocalbula ry of language } L_1 \text{ (Say English)}$$

$$V_F = \text{vocabular y of language } L_2 \text{ (Say Hindi)}$$

E1 *what is in a name ?*

नाम में क्या है ?

naam meM kya hai ?

F1 *name in what is ?*

E2 *That which we call rose, by any other name will smell as sweet.*

जिसे हम गुलाब कहते हैं, और भी किसी नाम से उसकी कुशबू समान मीठा होगी

F2 *Jise hum gulab kahte hai, aur bhi kisi naam se uski khushbu samaan mitha hogii*

That which we rose say , any other name by its smell as sweet

That which we call rose, by any other name will smell as sweet.

Vocabulary mapping

Vocabulary

V_E	V_F
what , is , in, a , name , that, which, we , call ,rose, by, any, other, will, smell, as, sweet	*naam, meM, kya, hai, jise, ham, gulab, kahte, aur, bhi, kisi, bhi, uski, khushbu, saman, mitha, hogii*

Key Notations

English vocabulary : V_E
French vocabulary : V_F
No. of observations / sentence pairs : S
Data D which consists of S observations looks like,

$$e^1{}_1, e^1{}_2, \dots, e^1{}_{l^1} \Leftrightarrow f^1{}_1, f^1{}_2, \dots, f^1{}_{m^1}$$

$$e^2{}_1, e^2{}_2, \dots, e^2{}_{l^2} \Leftrightarrow f^2{}_1, f^2{}_2, \dots, f^2{}_{m^2}$$

.....

$$e^s{}_1, e^s{}_2, \dots, e^s{}_{l^s} \Leftrightarrow f^s{}_1, f^s{}_2, \dots, f^s{}_{m^s}$$

.....

$$e^S{}_1, e^S{}_2, \dots, e^S{}_{l^S} \Leftrightarrow f^S{}_1, f^S{}_2, \dots, f^S{}_{m^S}$$

No. words on English side in s^{th} sentence : l^s
No. words on French side in s^{th} sentence : m^s
$index_E(e^s{}_p)$ =Index of English word $e^s{}_p$in English vocabulary/dictionary
$index_F(f^s{}_q)$ =Index of French word $f^s{}_q$in French vocabulary/dictionary

(Thanks to Sachin Pawar for helping with the maths formulae processing)

Hidden variables and parameters

Hidden Variables (Z) :
Total no. of hidden variables $= \sum_{s=1}^{S} l^s \, m^s$ where each hidden variable is as follows:
$z_{pq}^{s} = 1$, if in s^{th} sentence, p^{th} English word is mapped to q^{th} French word.
$z_{pq}^{s} = 0$, otherwise

Parameters (Θ) :
Total no. of parameters $= |\, V_E| \times |V_F|$, where each parameter is as follows:
$P_{i,j}$ = Probability that i^{th} word in English vocabulary is mapped to j^{th} word in French vocabulary

Likelihoods

Data Likelihood *L(D; Θ)* :

$$L(D;\Theta) = \prod_{s=1}^{S}\prod_{p=1}^{l^s}\prod_{q=1}^{m^s}\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)^{z_{pq}^s}$$

Data Log-Likelihood LL(D; Θ) :

$$LL(D;\Theta) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} z_{pq}^s \log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Expected value of Data Log-Likelihood E(LL(D; Θ)) :

$$E(LL(D;\Theta)) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} E(z_{pq}^s)\log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Constraint and Lagrangian

$$\sum_{j=1}^{|V_F|} P_{i,j} = 1 \ \ ,\forall i$$

$$\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} E(z_{pq}^s)\, log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right) - \sum_{i=1}^{|V_E|}\lambda_i\left(\sum_{j=1}^{|V_F|}P_{i,j} - 1\right)$$

Differentiating wrt P_{ij}

$$\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\left(\frac{E(z_{pq}^s)}{P_{i,j}}\right)-\lambda_i=0$$

$$P_{i,j}=\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)$$

$$\sum_{j=1}^{|V_F|}P_{i,j}=1=\sum_{j=1}^{|V_F|}\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)$$

Final E and M steps

M-step

$$P_{i,j} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{index_E(e_p^s),i}\, \delta_{index_F(f_q^s),j} E(z_{pq}^s)}{\sum_{j=1}^{|V_F|} \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{index_E(e_p^s),i}\, \delta_{index_F(f_q^s),j} E(z_{pq}^s)}, \forall i,j$$

E-step

$$E(z_{pq}^s) = \frac{P_{index_E(e_p^s),index_F(f_q^s)}}{\sum_{q'=1}^{m^s} P_{index_E(e_p^s),index_F(f_{q'}^s)}}, \forall s,p,q$$

PAN Indian SMT (whole word and subword)

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Pushpak Bhattacharyya, Shata-Anuvadak: Tackling Multiway Translation of Indian Languages, **LREC 2014**, Rekjyavik, Iceland, 26-31 May, 2014

Kunchukuttan & Bhattacharyya (EMNLP 2016)

Indian Language SMT (2014)

	hi	ur	pa	bn	gu	mr	kK	ta	te	ml	en
hi		61.28	68.21	34.96	51.31	39.12	37.81	14.43	21.38	10.98	29.23
ur	61.42		52.02	29.59	39.00	27.57	28.29	11.95	16.61	8.65	22.46
pa	73.31	56.00		29.89	43.85	30.87	30.72	10.75	18.81	9.11	23.83
bn	37.69	32.08	31.38		28.14	22.09	23.47	10.94	13.40	8.10	18.76
gu	55.66	44.12	45.14	28.50		32.06	30.48	12.57	17.22	8.01	19.78
mr	45.11	32.60	33.28	23.73	32.42		27.81	10.74	12.89	7.65	17.62
kK	41.92	34.00	34.31	24.59	31.07	27.52		10.36	14.80	7.89	17.07
ta	20.48	18.12	15.57	13.21	16.53	11.60	11.87		8.48	6.31	11.79
te	28.88	25.07	25.56	16.57	20.96	14.94	17.27	8.68		6.68	12.34
ml	14.74	13.39	12.97	10.67	9.76	8.39	9.18	5.90	5.94		8.61
en	28.94	22.96	22.33	15.33	15.44	12.11	13.66	6.43	6.55	4.65	

Baseline PBSMT - % BLEU scores (S1)

- **Clear partitioning of translation pairs by language family pairs**, based on translation accuracy.
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult

(Anoop Kunchukuttan, Abhijit Mishra, Pushpak Bhattacharyya, LREC 2014)

EBMT

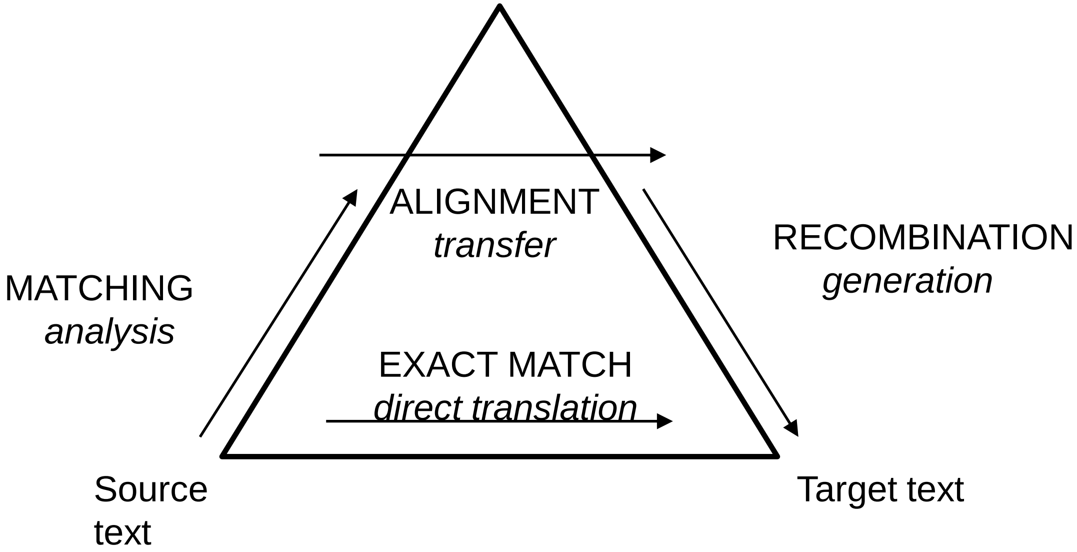
Nagao’s seminal paper 1984 (1/2)

“Man does not translate a simple sentence by doing ***deep linguistic analysis***, rather, man does the translation, first, by properly decomposing an input sentence into certain ***fragmental phrases*** (very often, into case frame units), and then

… (p.t.o)

Nagao’s seminal paper 1984 (2/2)

by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the ***analogy*** translation principle with proper examples as its reference”



The “Vauquois pyramid” adapted for EBMT

Analogy: the crux of the matter (need to emphasise)

- Needs measure of similarity
  - similar texts should indeed be *measured* as similar and dissimilar ones as dissimilar

- Means and Resources for measuring similarity.

Different ways of measuring text similarity

- Bag of words (BoW) based
- Permutation based
- N-gram based
- Vector based
- Tree based
- Semantic graph based
- Feature based

N-gram based matching: BLEU score

Recall -> Brevity Penalty

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

Precision -> Modified n-gram precision

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

$$\text{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right)$$

C: candidate sentence(s); C’: reference sentence(s); clip: to clip the count to max number of occurrences of an n-gram in the corpus; wn: weightage to a particular n-gram precision

Feature based (very rich)

$$S(I,R) = \frac{\sum_{i=1}^{n} w_i \times s(f_i^I, f_i^R)}{\sum_{i=1}^{n} w_i}$$

Sl. No.	Feature	Value	Similarity function *s(.)*
1	Length	Integer	equality
2	Active/Passive	1 (active)/ 0 (passive)	equality
3	Parse tree	--	Tree similarity between two parse trees
4	Concatenation of vectors of words forming the sentence	Vector of Boolean/real values	Cosine similarity
5	Bag of words forming the sentence	Set	Dice/Jackard and such other similarity measures
6	Position of nouns of the sentence in the wordnet hypernymy hierarchy	A function combining the *information content* of the individual nouns	equality
7	Position of the two main verbs of the sentence in Verb Ocean[2]	"distance" between the two main verbs in Verb Ocean	A rule that says *similar* or *dissimilar*, depending on the distance being within a threshold or not
8	main verb, its type and argument frame as given by the verbnet[3], types of nouns semantically related to it	A slot-filler structure for each sentence	Equality or subset-check on the slots and their fillers
9	Frame semantic representation of the sentence as per Framenet[4]	Slot-filler structure	Equality or subset-check on the slots and their fillers

EBMT’s ‘decoding’: RECOMBINATION

- Null Adaptation
- Re-instantiation
- Abstraction and re-specialization
- Case based substitution
- Semantic graph or graph-part substitution

Example of re-instantiation

- Input: *Tomorrow, today will be yesterday*
- Example matched: *Yesterday, today was tomorrow*
- कल, आज कल था
- kal, aaj kal thaa
- Yesterday, today tomorrow was

(*kal* is ambiguous in Hindi standing for *both* ‘yesterday’ and ‘tomorrow’)

Re-instantiation: adjustments (boundary friction problem)

- *Yesterday, today,* and *tomorrow* are all hyponyms of day.
- Main predicates in the example sentence and the input sentences *was* and *will be*.
- So, *adjusting* for the difference in predicates and matching the arguments, the translation is obtained as:

Re-instantiation leading to translation

- कल, आज कल होगा
- kal, aaj kal hogaa
- HG: Tomorrow, today yesterday will_be

Neural Machine Translation

Encoder-Decoder model

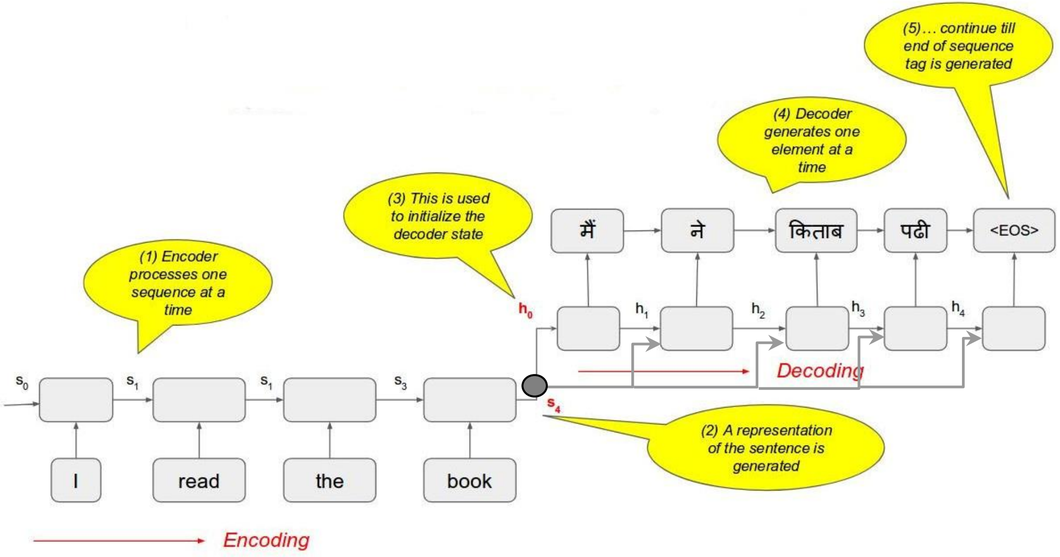


Image source- http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf

Some representative accuracy figure for Indian Language NMT

Language pair	**BLEU score**
Hi - Mr	31.25
Hi - Pa	63.38
Pa - Hi	68.31
Hi - Gu	49.98
Gu - Hi	**53.22** (↑ from 53.09 from SMT**)**

Comparing Knowledge based and data driven MT- with an example

Illustration of difference of RBMT, EBMT, SMT+NMT

- *Peter has a house*
- *Peter has a brother*
- *This hotel has a museum*

The tricky case of 'have' translation

English

- *Peter has a house*
- *Peter has a brother*
- *This hotel has a museum*

Marathi

- पीटरकडे एक घर आहे/ piitar kade ek ghar aahe
- पीटरला एक भाऊ आहे/ piitar laa ek bhaauu aahe
- ह्या हॉटेलमध्ये एक संग्रहालय आहे/ hyaa hotel madhye ek saMgrahaalay aahe

RBMT

If

syntactic subject is animate AND syntactic object is **owned** by subject

Then

“have” should translate to “kade … aahe”

If

syntactic subject is animate AND syntactic object denotes **kinship**

with subject

Then

“have” should translate to “laa … aahe”

If

syntactic subject is **inanimate**

Then

“have” should translate to “madhye … aahe”

EBMT

X have Y →

X_kade Y aahe /

X_laa Y aahe /

X_madhye Y aahe

SMT

- *has a house ←→ kade ek ghar aahe*

  <cm> one house has
- *has a car ←→ kade ek gaadii aahe*

  <cm> one car has
- *has a brother ←→ laa ek bhaau aahe*

  <cm> one brother has
- *has a sister ←→ laa ek bahiin aahe*

  <cm> one sister has
- *hotel has ←→ hotel madhye aahe*

  hotel <cm> has
- *hospital has ←→ haspital madhye aahe*

  hospital <cm> has

SMT: new sentence

"This hospital has 100 beds"

- *n*-grams (*n=1, 2, 3, 4, 5*) like the following will be formed:
  - *"This", "hospital",… (unigrams)*
  - *"This hospital", "hospital has", "has 100",… (bigrams)*
  - *"This hospital has", "hospital has 100", … (trigrams)*

DECODING !!!

IL-NLP: Challenges

Challenges of IL Computing (1/2)

- **Scale and Diversity: 22** major languages in India, written in 13 different scripts, with over 720 dialects

- **Code Mixing** ("kyo ye hesitation?")**; Gerundification** ("gaadi chalaaoing")

- **Absence of basic NLP tools and resources:** ref nlp pipeline

- **Absence of linguistic tradition for many languages**

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga and Ranjiva Munasinghe, *Indic Language Computing*, **CACM**, V 62(11), November 2019.

ILT Challenges (2/2)

- **Script complexity and non-standard input mechanism**: InScript Non-optimal
- **Non-standard transliteration** (“mango”→ ‘am”, “aam”, Am”)
- **Non-standard storage**: proprietary fonts
- **Challenging language phenomena**: Compound verbs (“has padaa”), morph stacking (“gharaasamorchyaanii”)
- **Resource Scarcity**

Mitigating the Resource problem

Three ways (1/2)

(1) Artificially boost the resource

- Subword based NLP
  - Characters, Syllables, Orthographic Syllables, Byte Pair Encoding
    - *Given, "khaa+uMgaa → will+eat"* AND *"jaa+rahaa_hE → is+going"*
    - *Produce "khaa+rahaa_hE → is+eatin*

Three ways (2/2)

(2) Take help from another language

–Cooperative NLP

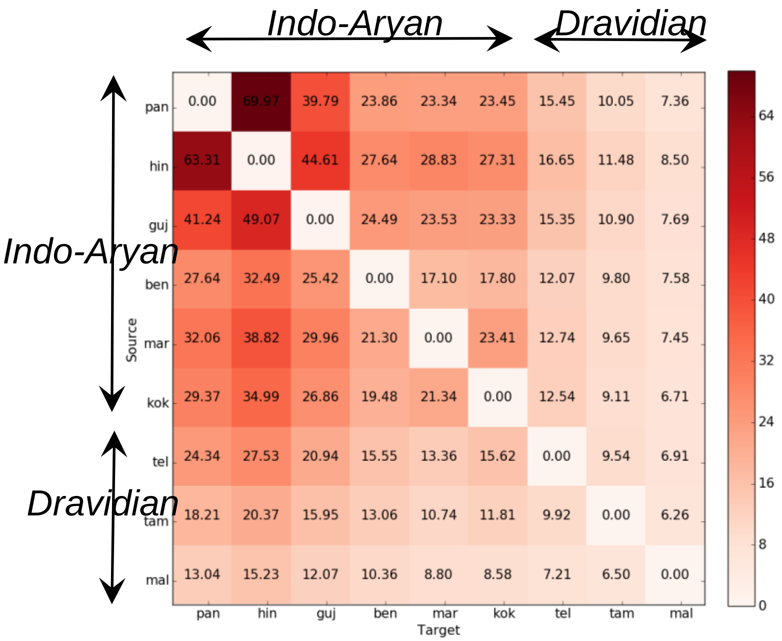
(3) Use “higher level language properties”

e.g., Part of Speech, Sense ID etc.

But there is a pitfall- NLP's "Law of Trade off"

- Trade Off:
  - *Precision vs. Recall*
  - Sparsity *vs.* Ambiguity
  - Information_Injection *vs.* Topic_Drift

Word level translation (BLEU scores)



Clear Partitioning based on language families

Translation between Indo Aryan languages is easiest

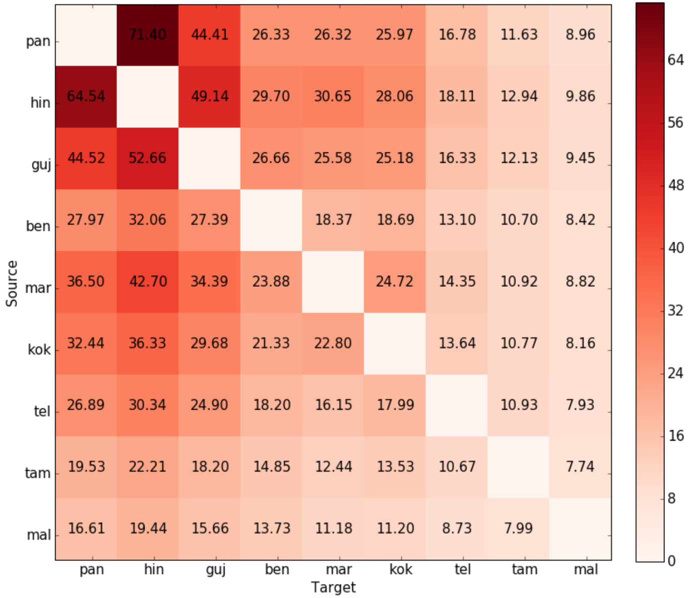
Translation into Dravidian languages is particularly difficult

Methods of sub-wording

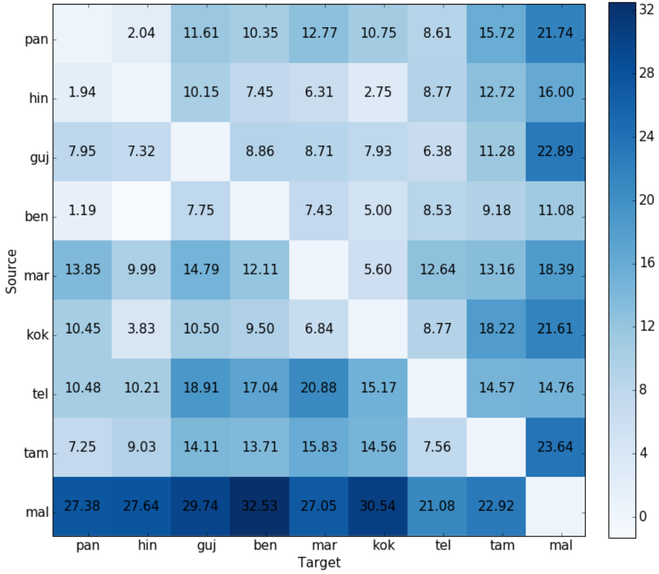
Subwords (for “jaauMgaa”)

- Characters: “j+aa+u+M+g+aa”
- Morphemes: “jaa”+”uMgaa”
- Syllables: “jaa”+”uM”+”gaa”
- Orthographic syllables: “jaau”+”Mgaa”
- BPE (depends on corpora, statistically frequent patterns): both “jaa” and “uMgaa” are likely

Morph level translation

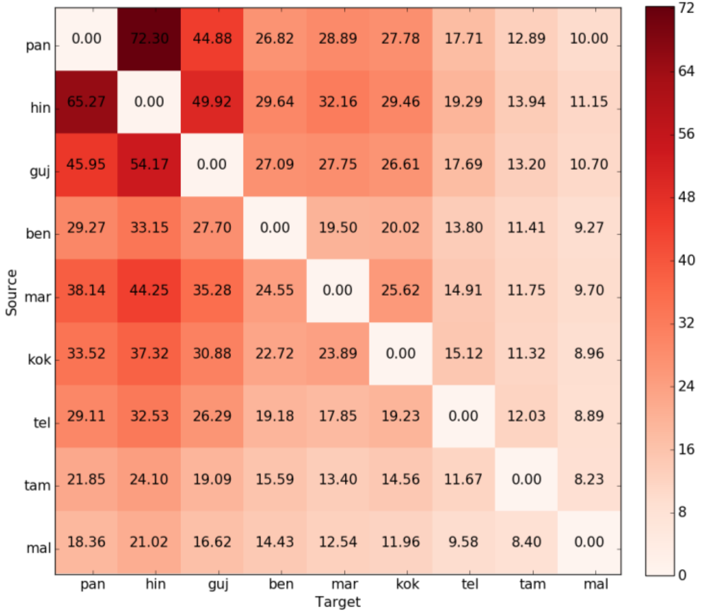


BLEU scores

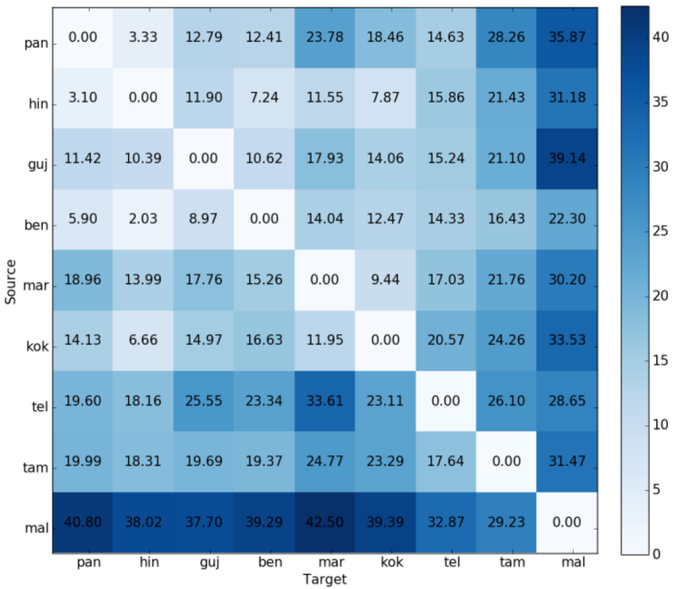


% improvement over word level scores

BPE level translation



BLEU scores



% improvement over word level scores

Factor based SMT

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh and Pushpak Bhattacharyya, *Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, **ACL-IJCNLP 2009**, Singapore, August, 2009.*

Semantic relations+Suffixes→Case Markers+inflections

I ate mangoes

↓

I {<agt} ate {eat@past} mangoes {<obj}

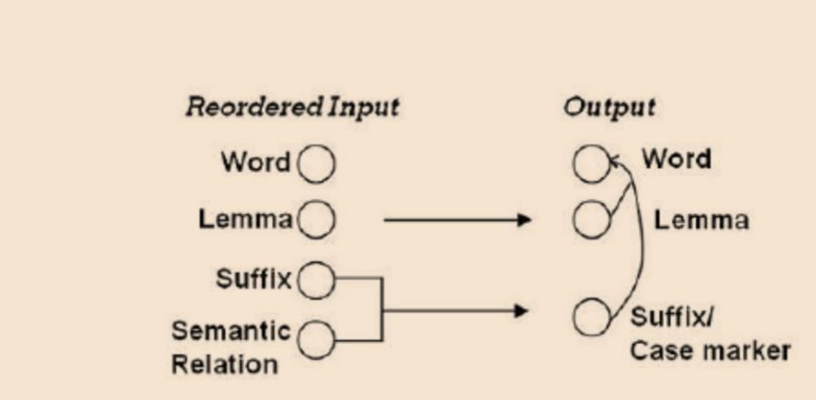
↓

I {<agt} mangoes {<obj.@pl} {eat@past}

↓

mei_ne aam khaa_yaa

Our Factorization based on Koehn and Hoang (2007)



1. a lemma to lemma translation factor (boy → लडक् (*ladak*))

2. a suffix + semantic relation to suffix/case marker factor (-s + subj → ए (*e*))

3. a lemma + suffix to surface form generation factor (लडक् + ए (*ladak* + *e*) → लडके (*ladake*))

Experiment: Corpus Statistics

	#sentences	#words
Training	12868	316508
Tuning	600	15279
Test	400	8557

Results: The impact of suffix and semantic factors

Model	BLEU	NIST
Baseline (surface)	24.32	5.85
lemma + suffix	25.16	5.87
lemma + suffix + unl	27.79	6.05
lemma + suffix + stanford	28.21	5.99

Results: The impact of reordering and semantic relations

Model	Reordering	BLEU	NIST
surface	distortion	24.42	5.85
surface	lexicalized	28.75	6.19
surface	syntactic	31.57	6.40
lemma + suffix + stanford	syntactic	31.49	6.34

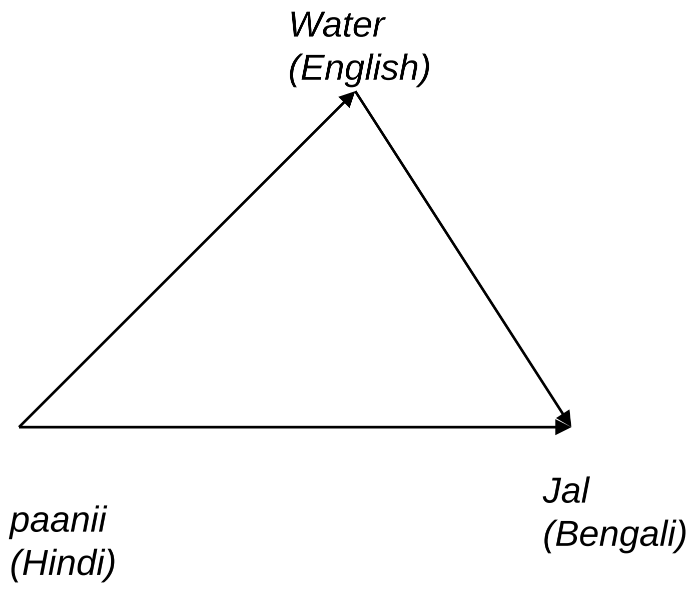
Subjective Evaluation: The impact of reordering and semantic relations

Model	Reordering	Fluency	Adequacy	#errors
surface	lexicalized	2.14	2.26	2.16
surface	syntactic	2.6	2.71	1.79
lemma + suffix + stanford	syntactic	2.88	2.82	1.44

Cooperative NLP: Pivot Based MT

Raj Dabre, Fabien Cromiere, Sadao Kurohash and Pushpak Bhattacharyya, *Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages*, **NAACL 2015**, Denver, Colorado, USA, May 31 - June 5, 2015.

Triangulation



L1→bridge→L2 *(Wu and Wang 2009)*

- Resource rich and resource poor language pairs
- Question-1: How about translating through a ‘bridge’?
- Question-2: how to choose the bridge?

Mathematical preliminaries

$$e_{best} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f})$$
$$= \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p_{LM}(\mathbf{e})$$

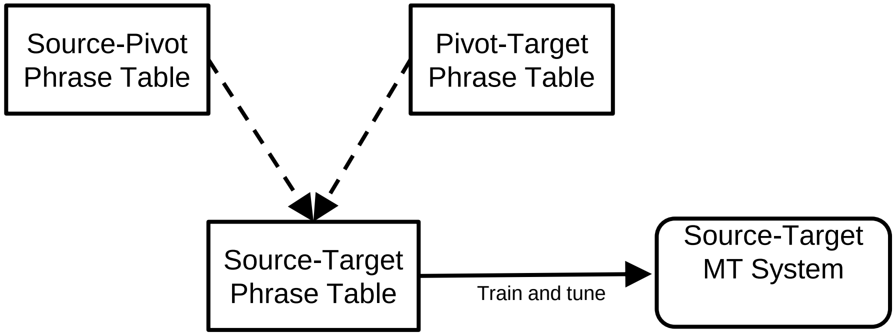
Where p($\boldsymbol{f}|\boldsymbol{e}$) is given by:

$$p(\mathbf{f}|\mathbf{e}) = p\left(\bar{f}^I|\bar{e}^I\right) = \prod_{i=1}^{I} \emptyset\left(\bar{f}_i|\bar{e}_i\right) d\left(a_i - b_{i-1}\right) p_w\left(\bar{f}_i|\bar{e}_i, a\right)^{\gamma}$$

$$\emptyset\left(\bar{f}_i|\bar{e}_i\right) = \sum_{\bar{p}_i} \emptyset\left(\bar{f}_i|\bar{p}_i\right) \emptyset\left(\bar{p}_i|\bar{e}_i\right)$$

$$p_w\left(\bar{f}_i|\bar{e}_i, a\right) = \prod_{l=1}^{n} \frac{1}{|m/(l,m) \in a|} \sum_{\forall (l,m) \in a} w(f_l/e_l)$$

Triangulation approach



- Important to induce language dependent components such as phrase translation probability and lexical weight

Mauritian Creole (MCR) →French (FR) →English (E)

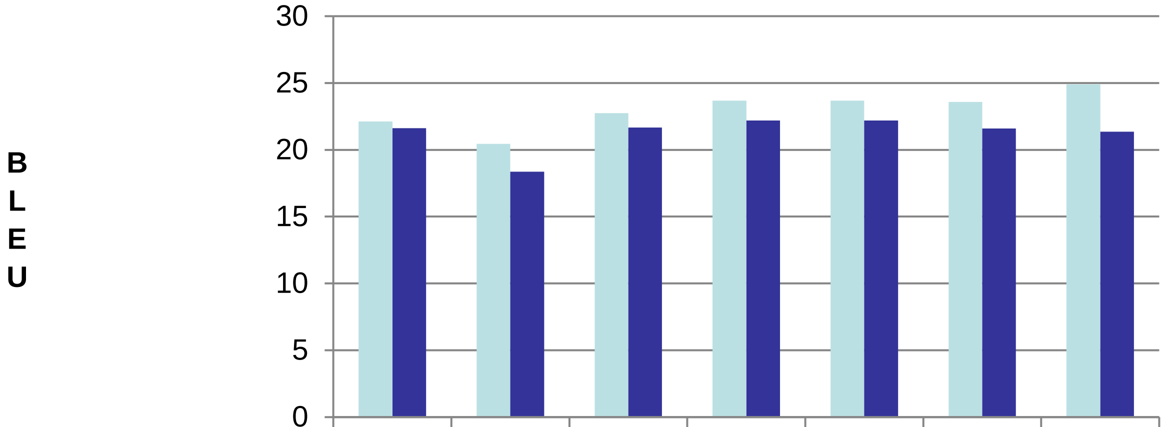
- MCR and FR share vocabulary and structure

French	Creole	English
avion	Avion	aeroplane
bon	Bon	good
gaz	Gaz	gas
bref	bref	brief
pion	pion	pawn

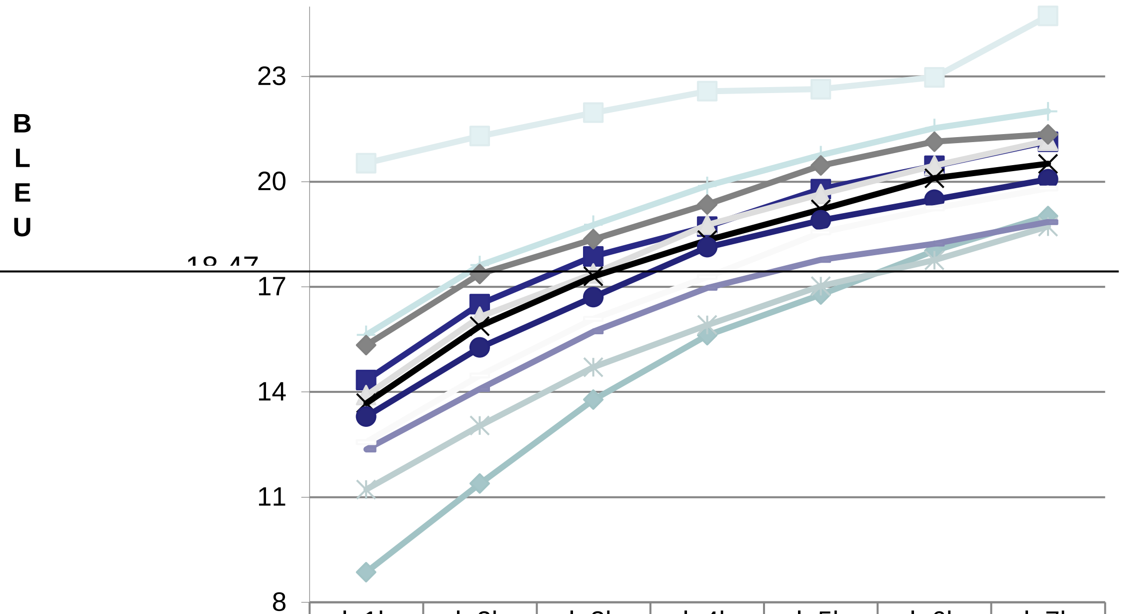
Experiment on MCR→FR→E

Language pair	#Sentences	#unique words (L1-L2)
En-Fr	2000000	127405- 147812
En-Cr (train + tune)	25010	16294-17389
En-Cr (test)	284 (142 short + 142 long)	1168-1070 + 3562-3326
Fr-Cr	18354	13769-13725

Results



	CR-EN-BASIC	CR-EN-MOD	CR-EN-COMBO	CR-EN-BACKOFF-1	CR-EN-BACKOFF-2	CR-EN-BACKOFF-3	CR-EN-BACKOFF-4
Short-Sentences	22.13	20.44	22.74	23.69	23.69	23.58	24.9
Long Sentences	21.61	18.37	21.66	22.2	22.2	21.59	21.36



	l=1k	l=2k	l=3k	l=4k	l=5k	l=6k	l=7k
DIRECT_l	8.86	11.39	13.78	15.62	16.78	18.03	19.02
DIRECT_l+BRIDGE_BN	14.34	16.51	17.87	18.72	19.79	20.45	21.14
DIRECT_l+BRIDGE_GU	13.91	16.15	17.38	18.77	19.65	20.46	21.17
DIRECT_l+BRIDGE_KK	13.68	15.88	17.3	18.33	19.21	20.1	20.51
DIRECT_l+BRIDGE_ML	11.22	13.04	14.71	15.91	17.02	17.76	18.72
DIRECT_l+BRIDGE_MA	13.3	15.27	16.71	18.13	18.9	19.49	20.07
DIRECT_l+BRIDGE_PU	15.63	17.62	18.77	19.88	20.76	21.53	22.01
DIRECT_l+BRIDGE_TA	12.36	14.09	15.73	16.97	17.77	18.23	18.85
DIRECT_l+BRIDGE_TE	12.57	14.47	16.09	17.28	18.55	19.24	19.81
DIRECT_l+BRIDGE_UR	15.34	17.37	18.36	19.35	20.46	21.14	21.35
DIRECT_l+BRIDGE_PU_UR	20.53	21.3	21.97	22.58	22.64	22.98	24.73

link

Neural ILMT

NMT with embellishments *(Minor revision, Journal of Machine Translation)*

- **Phrase table injection (PTI)**: supplying ‘good’ phrases from SMT system as additional data source to NMT system.
- **Word as feature:** merging word along with BPE segment to mitigate context loss.
- **Morph-seg-word:** morpheme segmentation followed by BPE, and then merging original morpheme and word to BPE segment.
- We report results for 56 systems for each of the above techniques.

Neural MT (NMT)

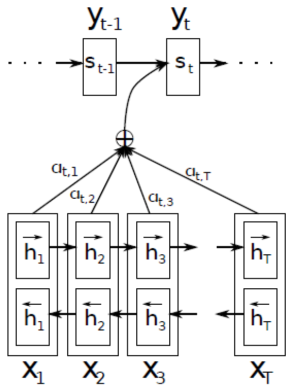
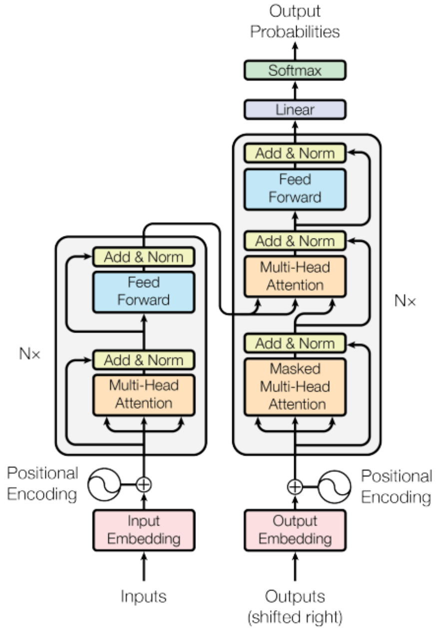


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

BiLSTM encoder decoder [3]



Transformer [8]

Language independent NMT

- **Languages chosen:**

Language	Hindi	Punjabi	Bengali	Gujarati	Marathi	Tamil	Telugu	Malayalam
Code	hi	pa	bn	gu	mr	ta	te	ml

Indo-Aryan (**IA**) family   Dravidian (**DR**) family

- **Model**: BiLSTM (details)
- **Dataset:** ILCI 1. Tourism and health domains. Dataset size in terms of number of sentences:

Training set	Tune set	Test set
46277	2000	500

En <-> {Mr, Hi} transformer output

- Dataset: ILCI1
- Explored lower range of merge operations.

	BLEU w/ BPE-0k	BLEU w/ BPE-2.5k	BLEU w/ BPE-5k	SMT
En-Mr	10.79	**14.26**	13.48	10.17
Mr-En	19.79	23.82	**24.19**	15.87
En-Hi	23.77	**29.18**	28.92	26.53
Hi-En	24.22	**31.22**	30.39	28.15

En <-> {Mr, Hi} Dataset: ILCI1 + PMIndia

Dataset size (no. of sentences):

	Train	Tune	Test
En-Hi	100267	1068	4273
En-Mr	80602	861	3445
Hi-Mr	92981	1000	4000

En <-> {Mr, Hi} Baselines on ILCI1+PMIndia

	NMT BPE-2.5k	**NMT BPE-5k**	**NMT BPE-7.5k**	**SMT**
En-Mr	**14.51**	15.04	**15.08**	10.51
Mr-En	23.76	**24.15**	24.13	16.6
En-Hi	27.05	27.72	**27.89**	20.75
Hi-En	30.96	**31.86**	30.45	24.05
Hi-Mr	27.25	**27.39**	-	24.38
Mr-Hi	37.39	**37.75**	-	34.31

En <-> {Mr, Hi} PTI results on ILCI1+PMIndia

	BPE-2.5k	BPE-5k	BPE-7.5k	BPE-10k	Best Baseline BLEU	Improvement
En-Mr	14.63	**15.97**	15.69	-	14.51	+1.46
Mr-En	23.08	25.22	**25.26**	25.03	24.15	+1.11
En-Hi	25.96	28.79	**29.28**	29.23	27.89	+1.39
Hi-En	29.94	33.6	33.93	**34.6**	31.86	+2.74
Hi-Mr	-	27.98	**28.49**	-	27.39	+0.59
Mr-Hi	-	38.94	**39.4**	-	37.75	+1.65

En <-> {Mr, Hi} PTI + Back Translation (BT)

	NMT BPE-5k	NMT BPE-7.5k	Improvement over PTI model
En-Mr	**16.73**	**-**	+0.76
Mr-En	-	26.24	+0.98
En-Hi	-	30.08	+0.8
Hi-En	-	35.03	+1.1

En <-> {Mr, Hi} PTI + Forward Translation (FT)

	NMT BPE-5k	NMT BPE-7.5k	Improvement over PTI model
En-Mr	16.47	-	+0.5
Mr-En	-	25.9	+0.64
En-Hi	-	29.77	+0.49
Hi-En	-	34.48	+0.52

En <-> {Mr, Hi} PTI results

	NMT BPE-5k	NMT BPE-7.5k	NMT BPE-10k	Best Baseline BLEU	Improvement
En-Mr	**21.05**	-	-	20.64	+0.41
Mr-En	28.76	-	-	28.64	+0.12
En-Hi	-	-	34.32	35.17	-0.85
Hi-En	-	-	38.65	36.57	+2.08
Hi-Mr	-	29.09	-	28.67	+0.42
Mr-Hi	-	36.98	-	36.59	+0.39

En-Hi-Mr NMT with embellishments (consolidated)

Approach	**Hi-En**	**En-Hi**	**Hi-Mr**	**Mr-Hi**	**En-Mr**	**Mr-En**
C: Agnostic Training	41.9	37.95	29.1	37.08	**25.34**	**32.45**
D: PTI(Phrase Table Injection)	38.65	34.32	29.09	36.98	21.05	28.76
E: D + Enhancement	42.15	36.78	**29.18**	**37.23**	**21.91**	29.57
F: BERT augmented NMT	29.74	25.89	28.64	34.21	14.98	18.37
G: BPE+word(pretrained BPE embeddings)	49.65	43.05	25.40	31.45	-	-

Sample outputs for En-Mr (PTI + t12)

En-src: i do know some young persons , who are active in such campaigns .
Mr-ref: असे काही **युवक** मला माहीत आहेत जे अशा प्रकारची मोहीम चालवतात .
OP: असे काही **तरुण व्यक्ती** मला माहीत आहेत जे अशा प्रकारची **मोहीम चालवतात** .
GT: मला अशा काही **तरुण व्यक्ती** माहित आहेत, जे अशा **मोहिमांमध्ये सक्रिय आहेत**.
Bing: अशा मोहिमांमध्ये सक्रिय असलेले काही **तरुण** मला माहीत आहेत.

En-src: this day marks the birth anniversary of the iron man of india , sardar vallabhbhai patel , the unifying force in bonding us as a nation
Mr-ref: हा दिवस भारताचे लोहपुरुष सरदार वल्लभभाई पटेल यांच्या जयंतीचा **आहे जे देशाला ऐक्याच्या धाग्यात गुंफणारे महानायक होते** .
OP: आज आपल्या देशात लोहपुरुष सरदार वल्लभभाई पटेल यांची जयंती , **एक देश म्हणून एकत्र आणत आहोत .**
GT: हा दिवस भारतीय लोहपुरुष सरदार वल्लभभाई पटेल या जयंतीनिमित्त **आम्हाला राष्ट्र म्हणून जोडण्याचे एकत्रीकरण**
Bing: या दिवशी सरदार वल्लभभाई पटेल या लोहपुरुषाची जयंती आहे. (incomplete)

Hindi-Marathi Examples

Hi: यदि श्वास प्रणालिका में सूजन आ जाये तब भी रक्त मुँह के रास्ते बाहर आने लगता है ।

Reference Mr: जर श्वासनलिकेला सूज आली तरीही रक्त तोंडावाटे बाहेर येऊ लागते.

Model Op: जर श्वासनलिकेत सूज आली ***तरीदेखील*** रक्त तोंडावाटे बाहेर येऊ लागते .

Google: जरी श्वसन प्रणालीमध्ये सूज येत असेल तर, तोंडातून रक्त देखील बाहेर येते.

Hi: जब यह हिस्से तीव्रता से घटते हैं तो पेट थोड़ा भूखा रहता है और मस्तिष्क को भूख के संकेत देता है ।

Reference Mr: जेव्हा हे भाग वेगाना कमी होतात, तेव्हा पोट थोडेसे भूके राहतात.

Model Op: जेव्हा हा भाग तीव्रतेने कमी होत असतो तेव्हा पोट थोडे उपाशी राहतो आणि मस्तिष्काला भूकेचा संकेत देतो .

Google: जेव्हा हे भाग झपाट्याने कमी होतात तेव्हा पोट किंचित भूक राहते आणि मेंदूला उपासमारीचे संकेत देते.

Examples from Covid Domain

Hi: यदि स्वास्थ्य अनुमति देता है, तो नियमित रूप से घरेलू काम किया जाना चाहिए। पेशेवर काम को श्रेणीबद्ध तरीके से फिर से शुरू किया जाना है।

Model Op: जर आरोग्य परवानगी देवून माकडून तर नियमितपणे घरगुती काम केले जाणे अवर्णनीय निर्धाराला कठीण पद्धतीने पुन्हा सुरू केले जाणे

Google: आरोग्यास परवानगी मिळाल्यास घरातील कामे नियमितपणे करावीत. व्यावसायिक काम श्रेणीरित्या पुन्हा सुरू करावे लागेल.

Hi: रोज सुबह या शाम आराम से चलना जितना कि सहन किया जा सके ।

Model Op: रोज सकाळी किंवा संध्याकाळी आरामात चालणे जेवढे सहन केले जाऊ शकते .

Google: जितके सहन केले जाऊ शकते तितके दररोज सकाळी किंवा संध्याकाळी आरामात चालणे.

Examples from Programming Domain

Hi: दूसरी ओर एक व्हाइल लूप आम तौर पर इस्तेमाल किया जाता है जब आपको अग्रिम से नहीं पता होता है।

Model Op: दुसर्या बाजूला एक अपायकारक अनावश्यक वापर केला जातो जेव्हा तुम्हाला लगेच कळत नाही .

Google: दुसरीकडे जेव्हा आपल्याला आगाऊ माहिती नसते तेव्हा पांढरा पळवाट सामान्यत: वापरला जातो.

Hi: अब हम यह अंत से शुरू कर रहे हैं और केवल पहला कारक रख रहे है कि हम तो , तो हमने जो इस उदाहरण में देखा एक नए प्रकार का लूप है ।

Model Op: आता आम्ही हा शेवटपासून सुरू करत आहोत आणि केवळ पहिले कारण आहे की आपण तर या उदाहरणामध्ये पाहिले , एक नवीन प्रकारचा स्पष्ट आहे .

Google: आता आपण या टोकापासून सुरूवात करीत आहोत आणि फक्त पहिला घटक ठेवणे म्हणजे आपण, नंतर या उदाहरणात जे पाहिले ते एक नवीन प्रकारचे लूप आहे.

Disfluency Correction in the context of Speech to Speech MT *(under review for EACL 2021)*

- Pair: Disfluent English - Fluent English (Switchboard corpus)
- Domain: includes telephone conversations between strangers on specific topics.

Type	**Set**	**Disfluent Sentences**	**Fluent Sentences**
Non Parallel	Train	55,482	55,482
Parallel	Dev	11,889	11,889
	Test	11,889	11,889

Results

	Model	Validation	Test
Supervised	Sequence to Sequence (Bi-LSTM)	87.23	88.08
	BART	89.27	90.08
Unsupervised	Noise Induction (Transformer)	65.17	53.78
	Style Transfer (Bi-LSTM)	61.26	62.77
	Style Transfer (Transformer)	78.72	79.39
Semi-Supervised	Style Transfer (Transformer)	84.1	85.28

Semi-Supervised:

Amount of parallel data = 554 sentences (1% of train set)

Example Output

Type	Disfluent	BART	Seq-to-Seq	US(Bi-LSTM)	US(Transformer)	SS (Transformer)	Fluent
discourse, filler	so uh been a different turn	been a different turn	been a different turn	been a different turn	been a different turn	been a different turn	been a different turn
conjunction, repetition	but i i i find this whole	i find this whole	i find this whole	anyway i find it all	i find this whole	i find this whole	i find this whole
restart	it's you're you're taking words and developing a picture in your mind	you're taking words and developing a picture in your mind	you're taking words and developing a picture in your mind	it's you're taking chicken and tobacco words in a mind	it's taking words and developing and a picture in your mind	it's taking words and developing and a picture in your mind	you're taking words and developing a picture in your mind

US: Unsupervised, SS: Semi-supervised

Summary

- MT Paradigms
- Data Driven MT: SMT and NMT
- Tricks of Resource Mitigation
- Unsupervised NMT
- Experience of IL-NMT

Summary on resource mitigation tricks

- Several techniques explored and demonstrated their efficacy.
  - Phrase Table Injection, has great potential to boost BLEU scores, particularly when Dravidian languages are involved.
  - Harnessing monolingual data with back translation, forward translation is advantageous.
  - Enhancements like morph and word feature injection

Final Message

“NLP is a task in Trade Off”

e.g., Not too much of subwords or cooperation

(beware of ‘*ambiguity insertion*’),

not too little

(beware of ‘sparsity’) !!

“The middle path is the golden one”- Buddha

URLS

http://www.cse.iitb.ac.in/~pb

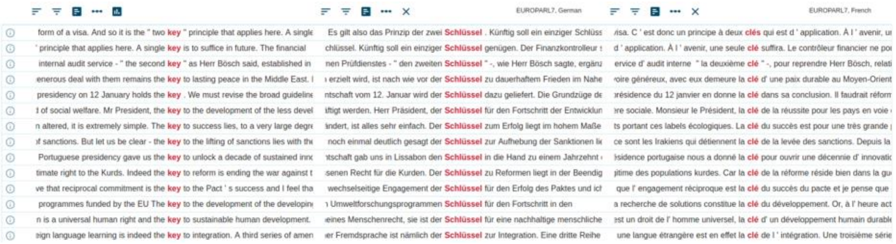
http://www.cfilt.iitb.ac.in

Thank You

Why is Unsupervised NMT needed?

Diptesh Kanoja

Unsupervised NMT - Why?



Supervised NMT

- Parallel Corpus
- Monolingual Corpus

Manual Translations → Cognitive Load

“Unsupervised” NMT

- **No parallel corpus**

However, the requirement is:

- **Large monolingual corpus**
- Cross-lingual Word Embeddings
- **Low-resource languages**

Image Source: Paramount Pictures

Resource Constraints

- Lack of resources for NLP tasks.
- Low resource languages.
  - Indian Languages including Sanskrit.
  - Hebrew, Greek, and Latin.
- Obscure Languages such as Sentinelese (North Sentinel Island, Indian Ocean), Ugaritic, etc.
- Monolingual corpus may be available.

Resource Generation/Building

- Parallel word mappings can be generated.
    - Unsupervised Embedding mappings (similar script).

- Word mappings can also be created manually.
    - For language written in different scripts, but human supervision is needed.

- Word representations form the crux of most NLP tasks.

Foundations

1. **Cross-lingual embeddings**
2. Denoising Autoencoder
3. Back-translation

Word Representation for Humans

In humans, the acquisition of information and creation of mental representations occurs in a two-step process. (Ramos et. al., 2014)

Sufficiently complex brain structure is necessary to establishing internal states capable to co-vary with external events.

The validity or meaning of these representations must be gradually achieved by confronting them with the environment.

Cross-lingual Word Embeddings

- The geometric relations that hold between words are similar across languages*.
    - For instance, numbers and animals in English show a similar (isomorphic) geometric structure as their Spanish counterparts.
- The vector space of a source languages can be transformed to the vector space of the target language t by learning a linear projection with a transformation matrix $W^{s\to t}$.

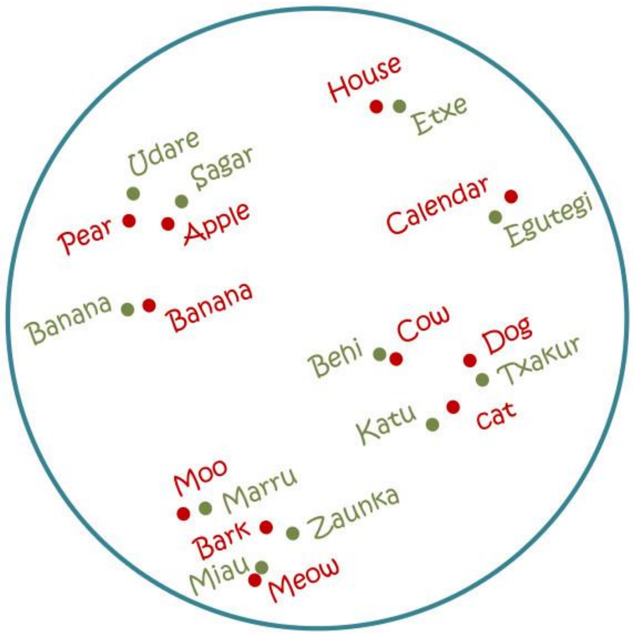
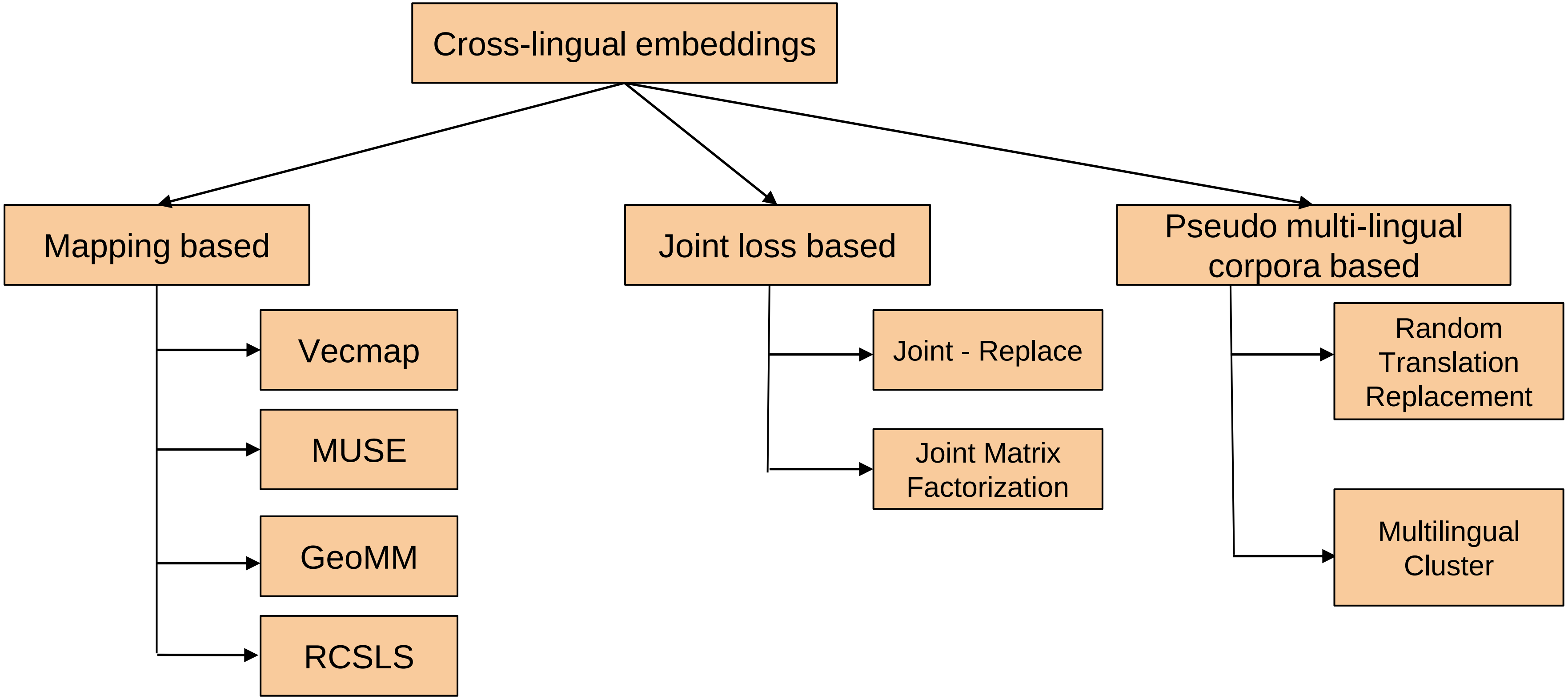
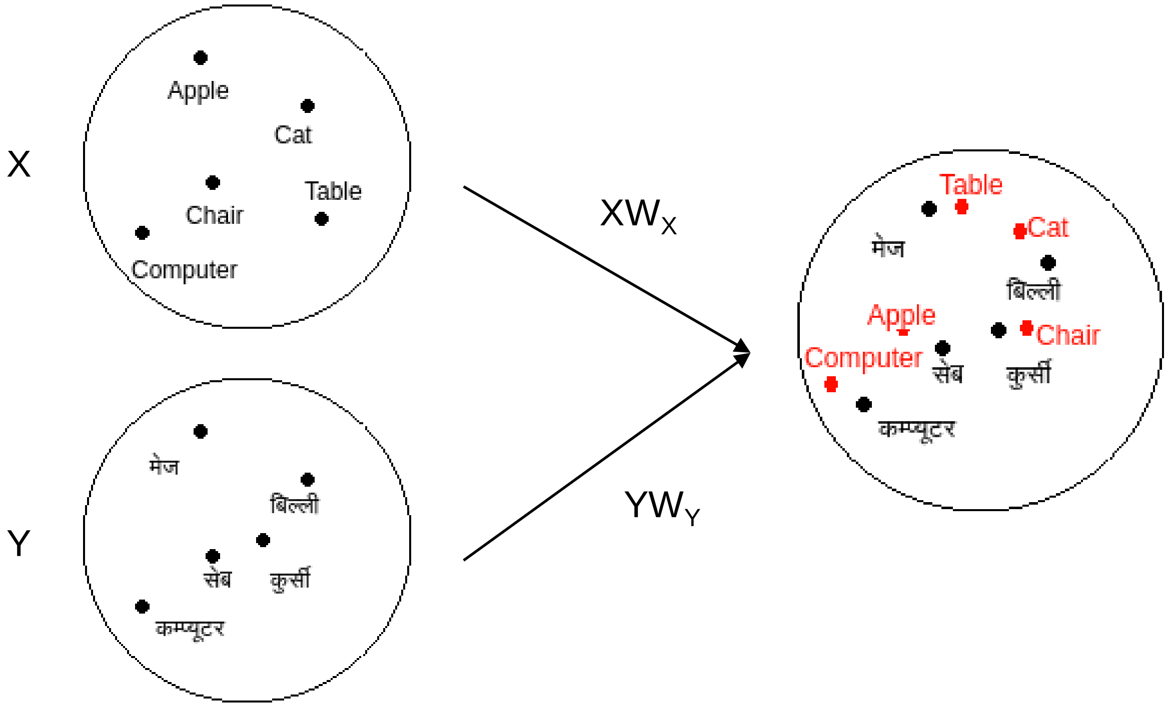


Image source- www.mikelartetxe.com

Cross-lingual embeddings: Approaches



Cross-lingual embeddings: Mapping based



- Task is to learn $\mathbf{W_X}$ and $\mathbf{W_Y}$ (the transformation matrices)
- X, Y are monolingual embedding spaces

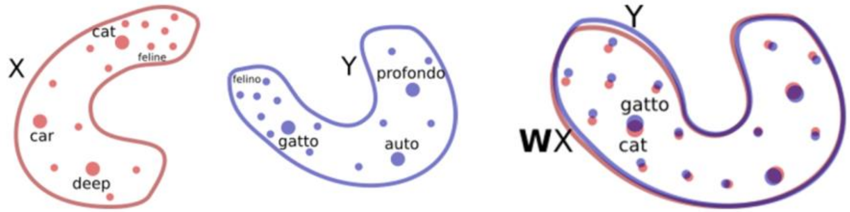
MUSE

Given, target Vector Y and source Vector X

Learns Mapping Y=XW.

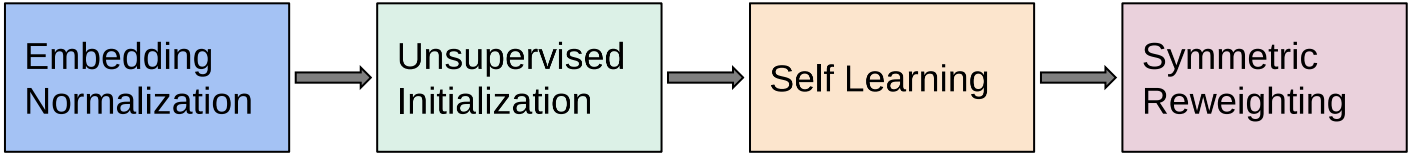
Trains a discriminator to tell whether two vectors are from the same language.

Also, a generator to map the vectors from one language into each other.



Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word translation without parallel data." *arXiv preprint arXiv:1710.04087* (2017).

VecMap (Artexe et al. 2018)



- Embeddings Normalization
  - Length normalization + Mean centering + Length normalization
- Unsupervised initialization
  - Assume both spaces are isometric
  - Nearest neighbor retrieval on XX^T and YY^T
- Self training
  - Compute the optimal orthogonal mapping by maximizing the similarity for the current dictionary D
  - Compute the dictionary over the similarity matrix of the mapped embeddings
- Symmetric weighting to induce good dictionary
  - $W_X = US^{1/2}$, $W_Y = VS^{1/2}$

Artetxe Mikel, Gorka Labaka, and Eneko Agirre. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings." *ACL 2018*.

Joint training + Cross-lingual alignment (Wang et al 2019)

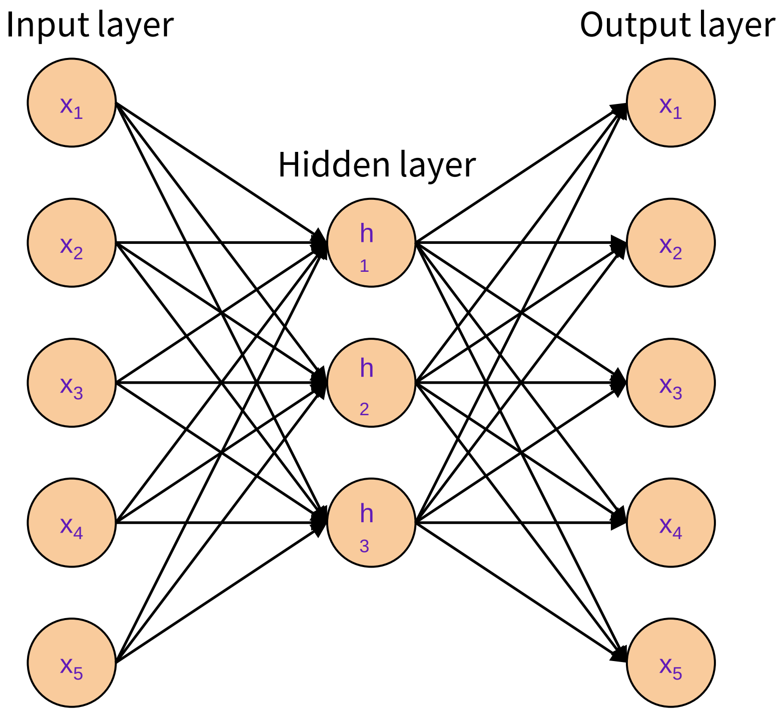
- Joint initialization
  - Joint training using monolingual embedding training algorithm using combined corpus
- Vocabulary reallocation
  - Create source, target and common vocabulary
- Alignment refinement
  - Mapping based algorithm for align source and target to the same space

Wang Z, Xie J, Xu R, Yang Y, Neubig G, Carbonell JG (2019) Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In: International Conference on Learning Representations

Foundations

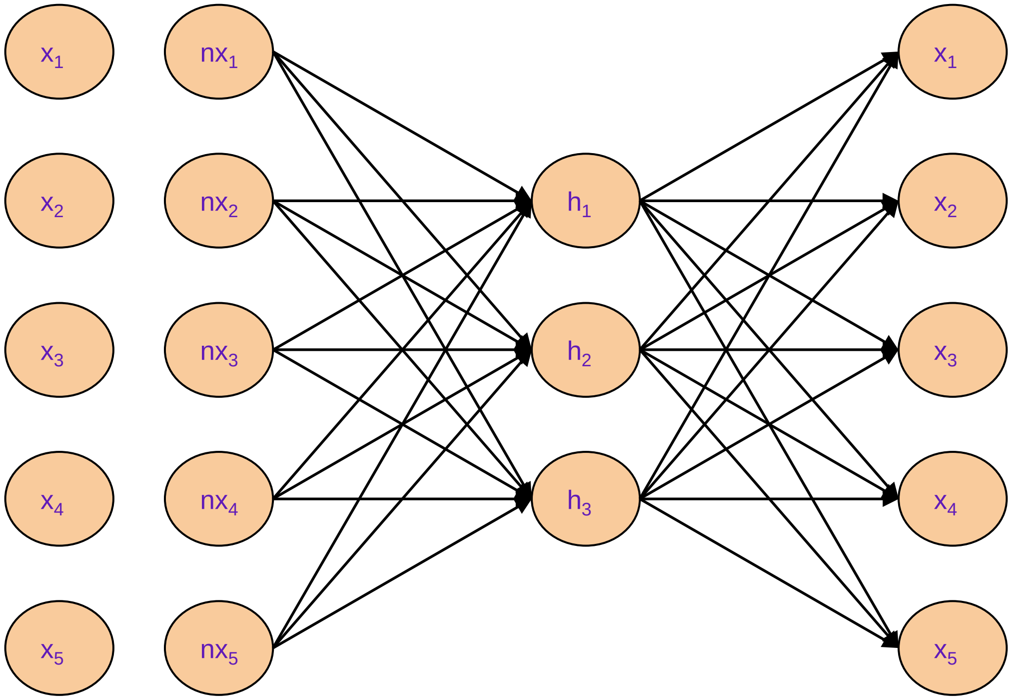
1. Cross-lingual embeddings
2. **Denoising Autoencoder**
3. Back-translation

Autoencoder



- Representation learning
- Neural network to learn reconstruction of the data
- Optimize **Reconstruction Error**
- Balance between
  - Accurately build a reconstruction
  - Handle inputs such that the model doesn't learn to copy the data

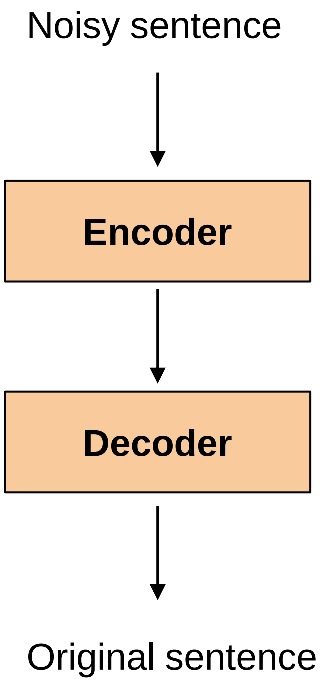
Denoising auto-encoder



Corrupted data

- Learn to generate original sentence from a noisy version of it
- Eliminates the learning of identity function

Denoising auto-encoder



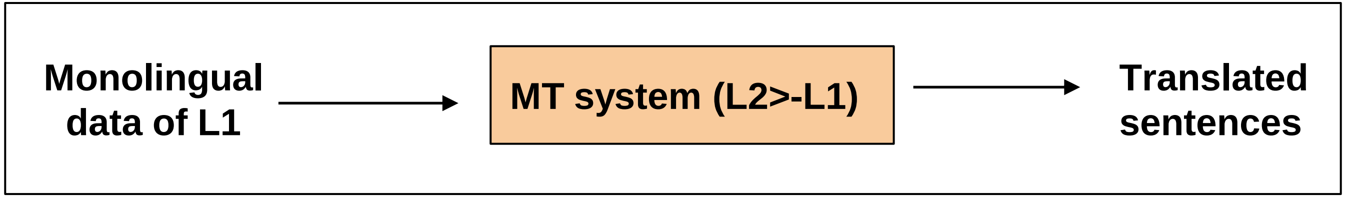
- Encoder representation is the representation for noisy sentence
- Decoder tries to generate the original sentence from the encoder representation of the noisy sentence
- A sentence can be corrupted using different types of noise
  - Swapping of words
  - Removal of words
  - Replacement of words with other words

Foundations

1. Cross-lingual embeddings
2. Denoising Autoencoder
3. **Back-translation**

Back-Translation

- Utilize monolingual data of target language
- Generate pseudo parallel data using MT system in opposite direction (target->source)

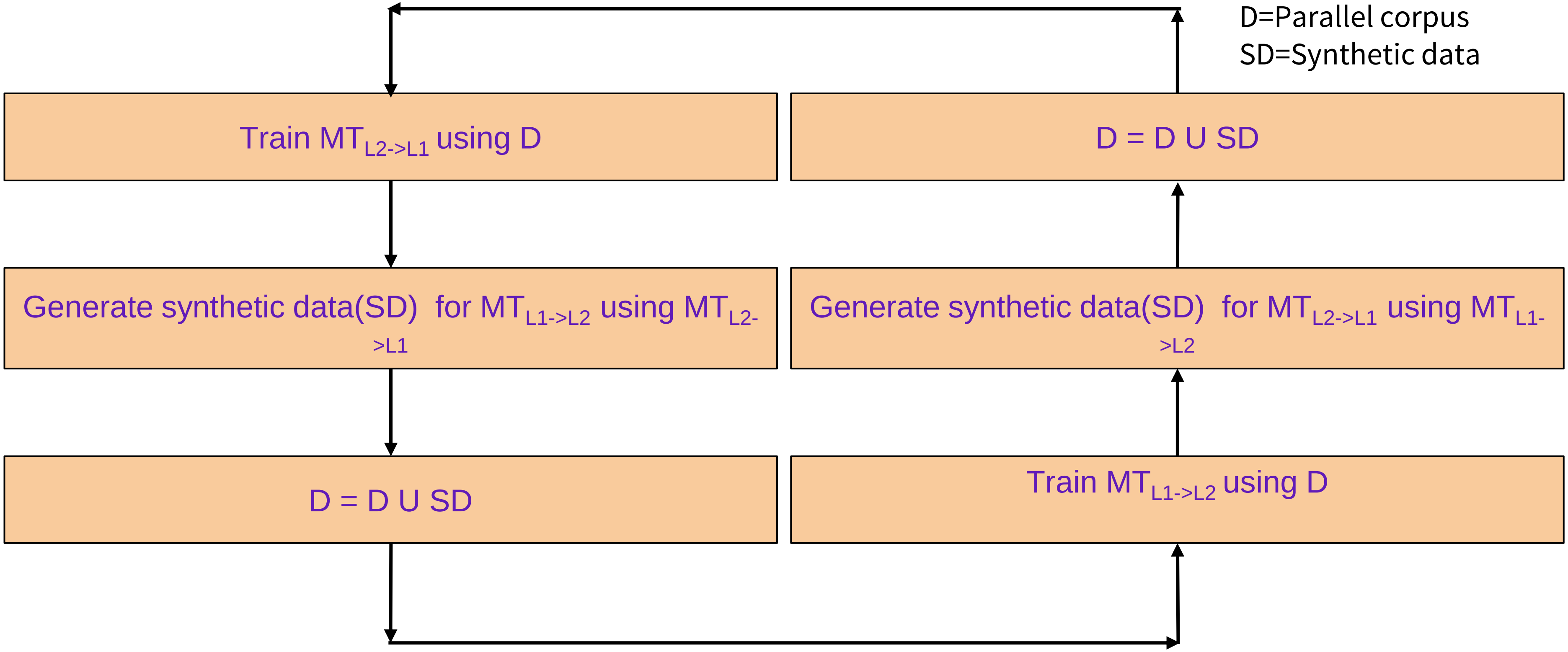


Generation of pseudo parallel sentences

- Train MT system (L1->L2) using a combination of parallel and generated synthetic data both

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

Iterative Back-Translation



Iterative Back-Translation

Setting	French–English		English–French		Farsi–English	English-Farsi
	100K	1M	100K	1M	100K	100K
NMT baseline	16.7	24.7	18.0	25.6	21.7	16.4
back-translation	22.1	27.8	21.5	27.0	22.1	16.7
back-translation iterative+1	22.5	-	22.7	-	22.7	17.1
back-translation iterative+2	22.6	-	22.6	-	22.6	17.2

- Beneficial for Low resource languages also

Image source: Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "Iterative back-translation for neural machine translation." In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.

UMT Approaches

Tamali Banerjee

1. **Unsupervised NMT**
2. GAN for UNMT
3. Unsupervised SMT
4. Hybrid UMT

Introduction

- In ICLR 2018, two concurrent papers showed that it is possible to train an NMT system without using any parallel data.

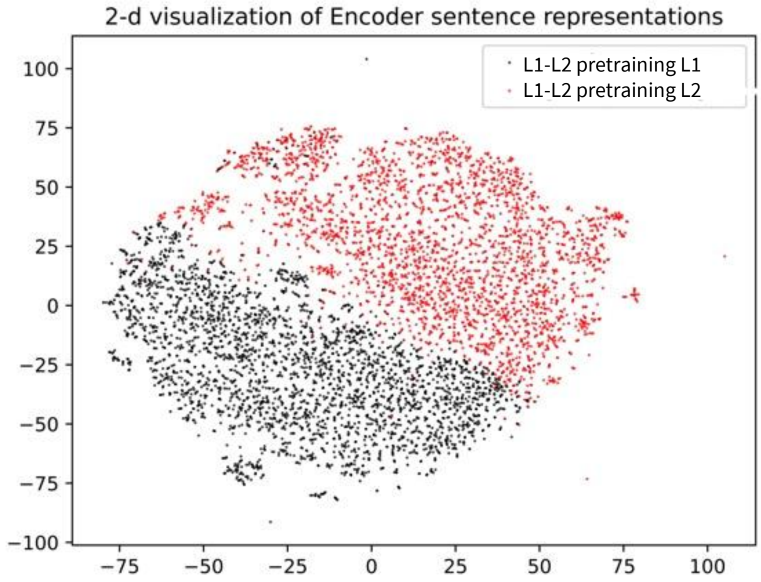
List of papers

1. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).
2. G. Lample, A. Conneau, L. Denoyer, MA. Ranzato. 2018. Unsupervised Machine Translation With Monolingual Data Only. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

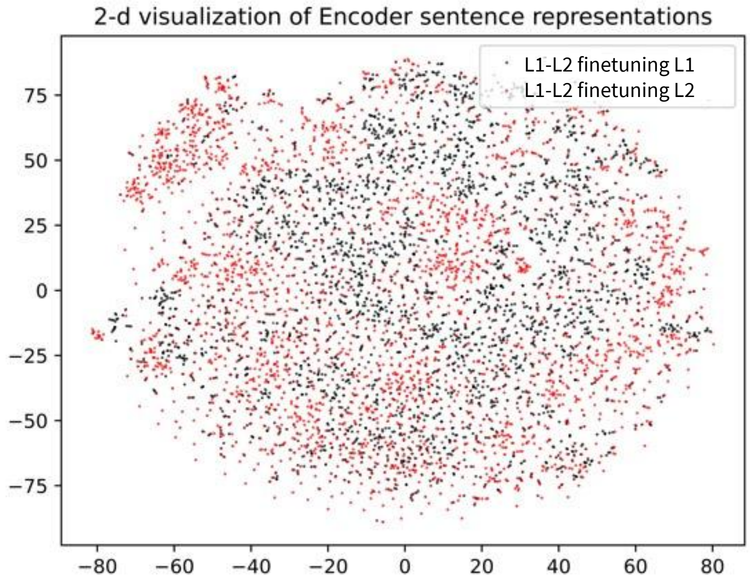
Components of U-NMT

- Bi-lingual embedding: It projects word embeddings of both languages in the same embedding space.
- Language modeling: It helps the model to encode and generate sentences.
  - Through initialization of the translation models.
  - Through iterative training.
- Iterative back-translation: It bridges the gap between encoder sentence representation in source and target languages.

Effect of Back-translation



Before Back-translation



After Back-translation

Image credit: Rudra and Jyotsana

Architecture

- Bi-lingual embedding layer
- Encoder-Decoder architecture
- Dual structure
- Sharing of modules

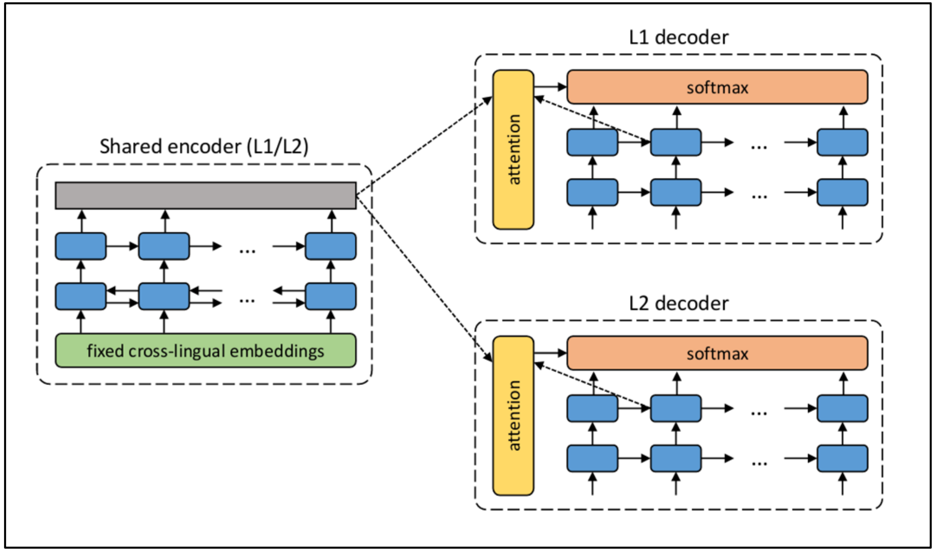
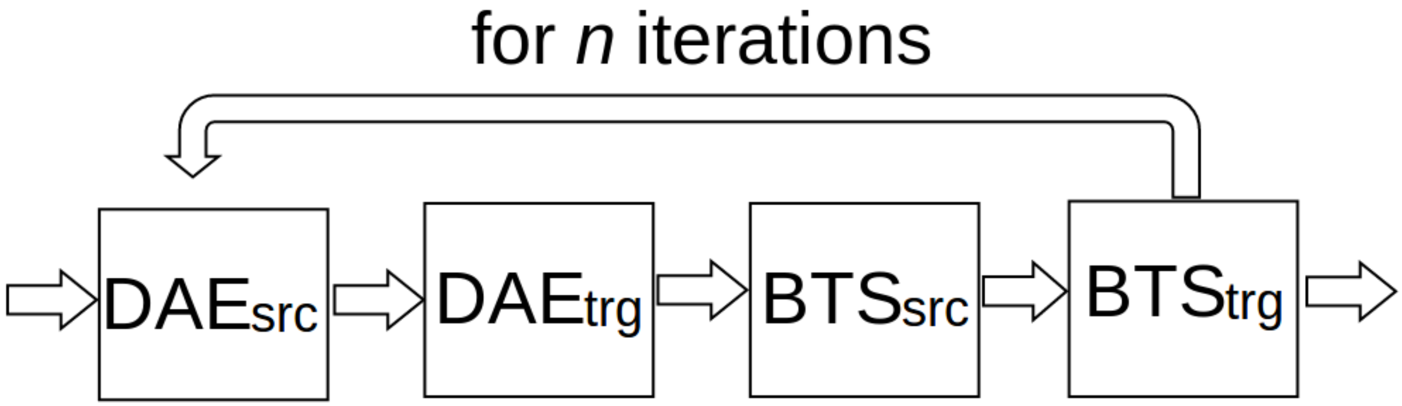


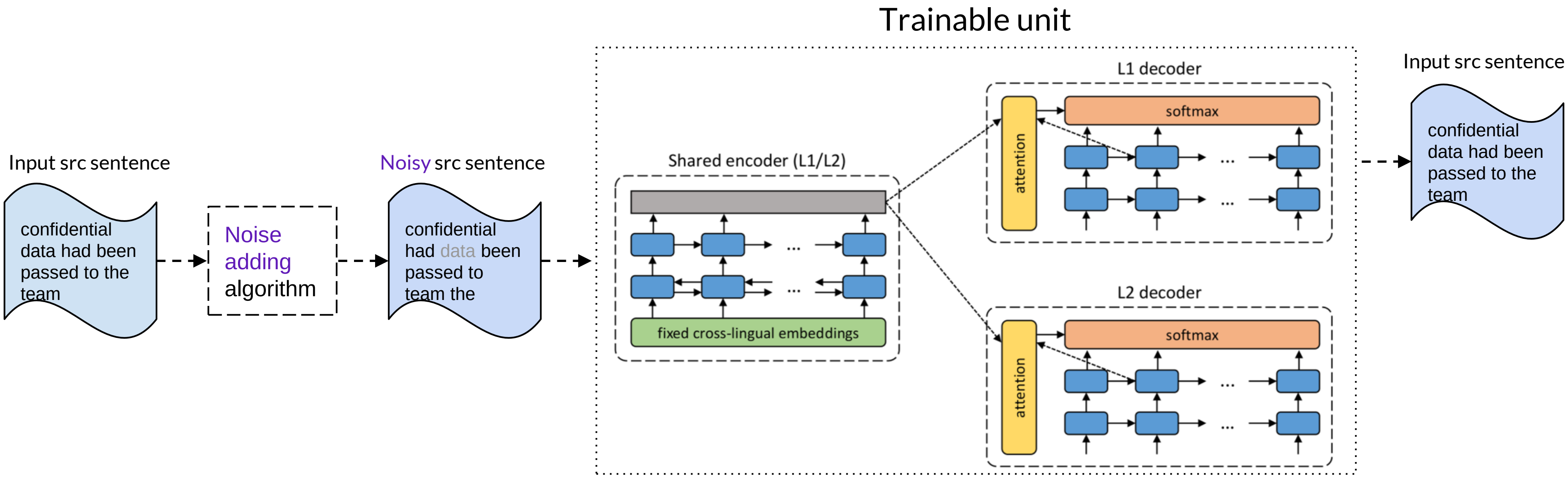
Image source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

Training Procedure



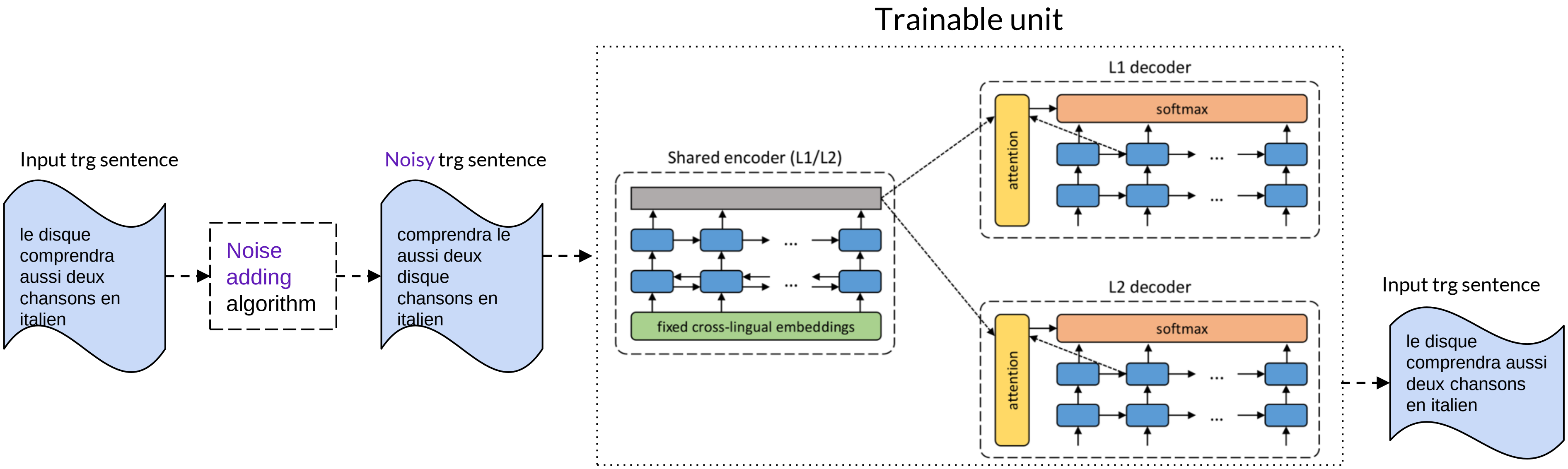
DAE_{src}: Denoising of source sentences; DAE_{trg}: Denoising of target sentences;
BTS_{src}: Back-translation with shuffled source sentences; BTS_{trg}: Back-translation with shuffled target sentences;
n : total number of iteration till it reaches stopping criterion.

U-NMT: Denoising of source sentences



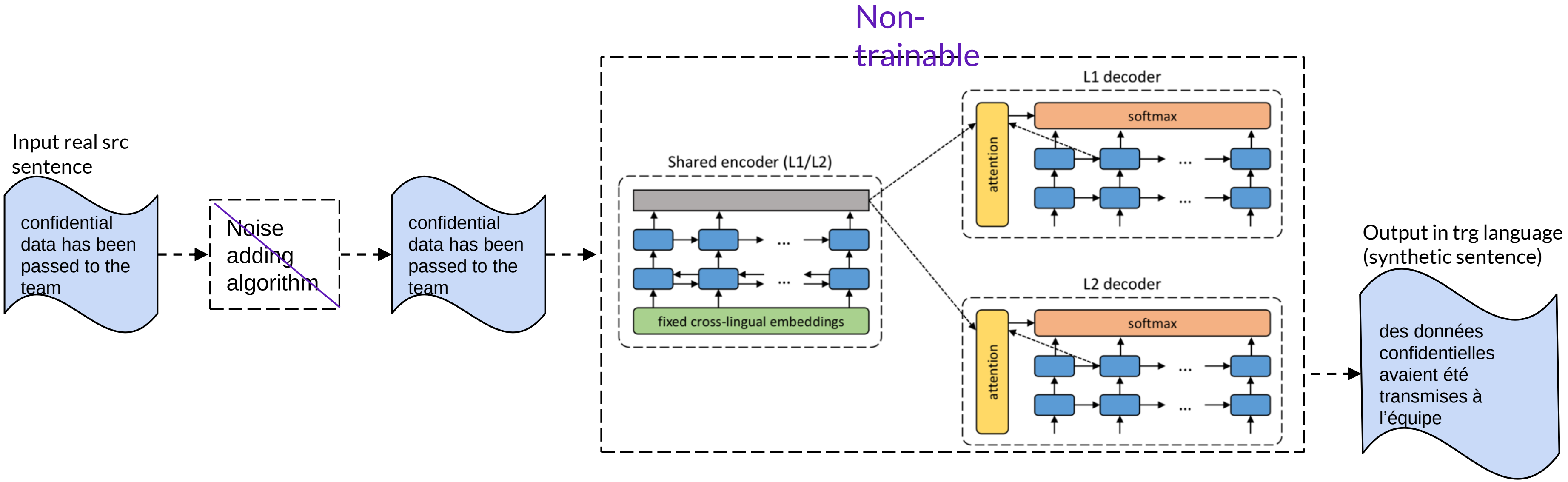
Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

U-NMT: Denoising of target sentences



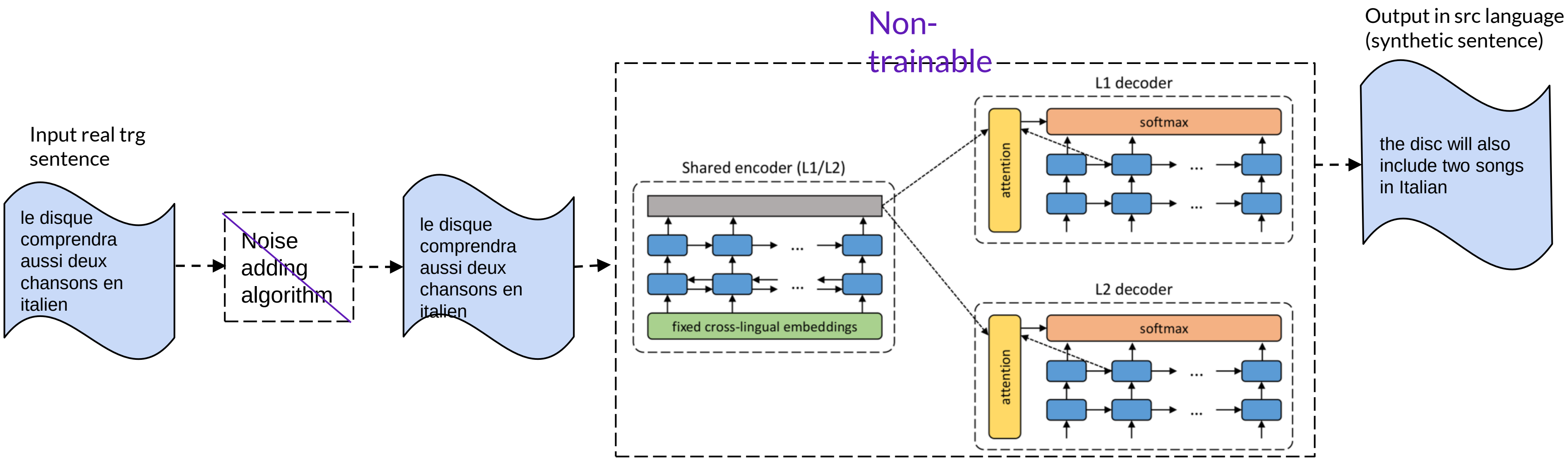
Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

U-NMT: Back-translation Corpus Construction (source to target)



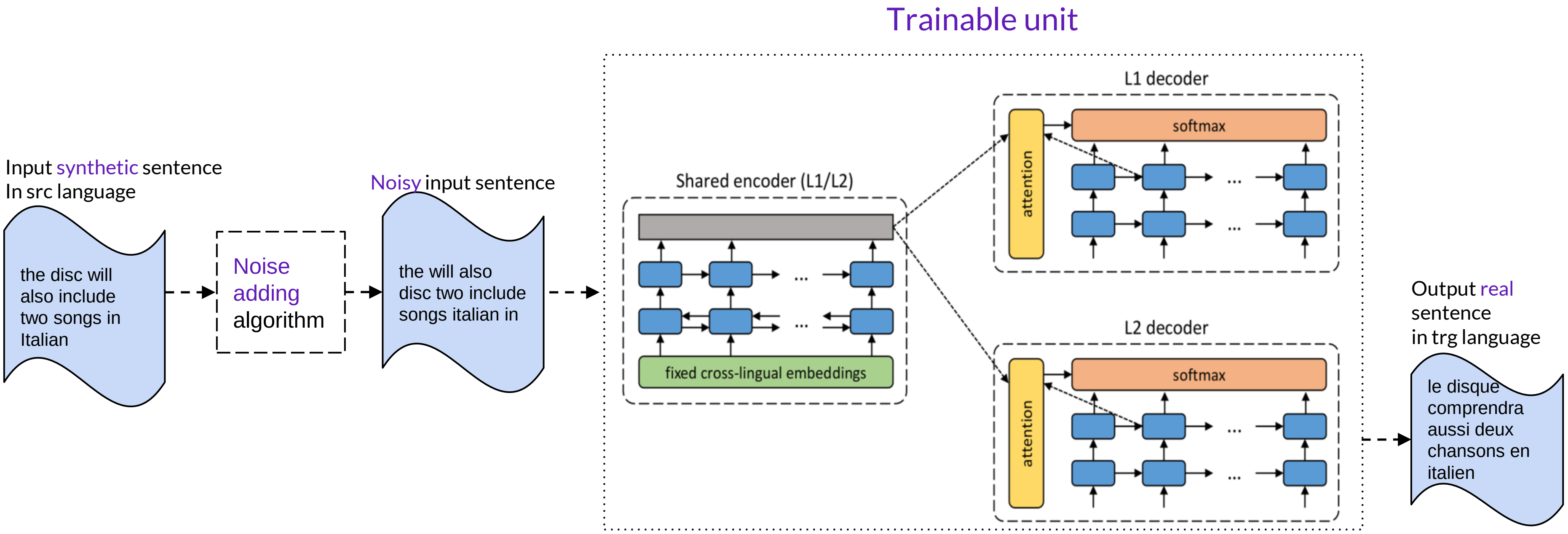
Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

U-NMT: Back-translation Corpus Construction (target to source)



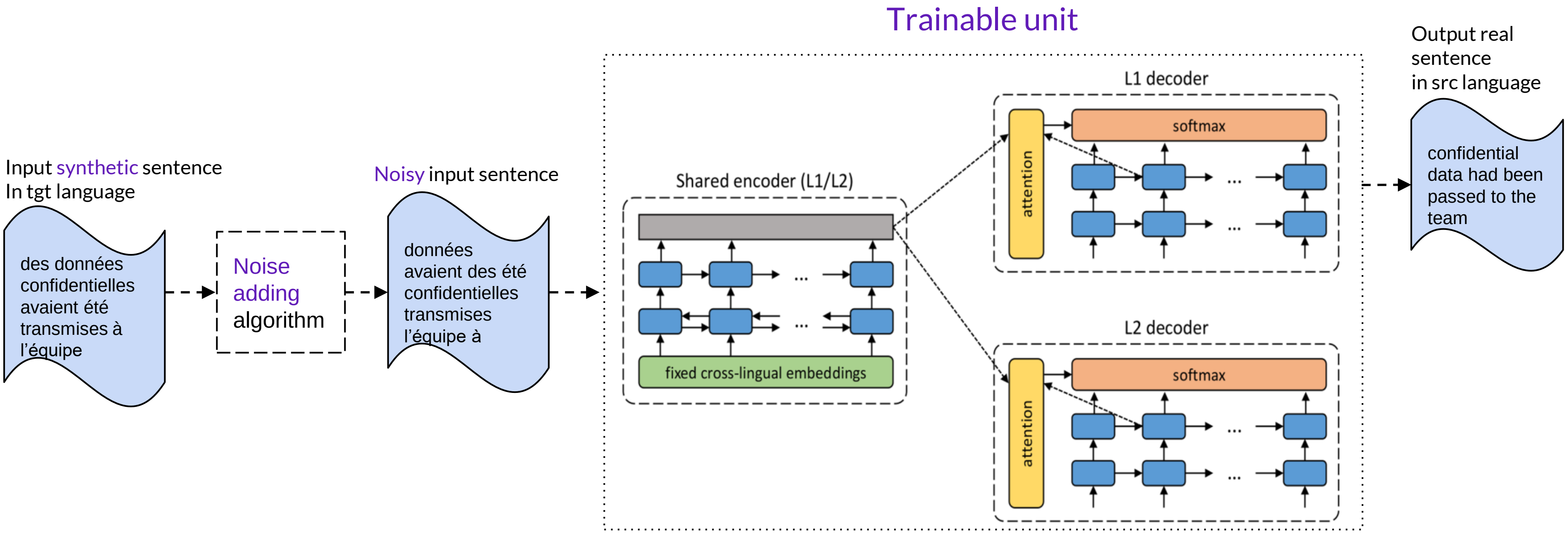
Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

U-NMT: Training with Back-translated data (source to target)



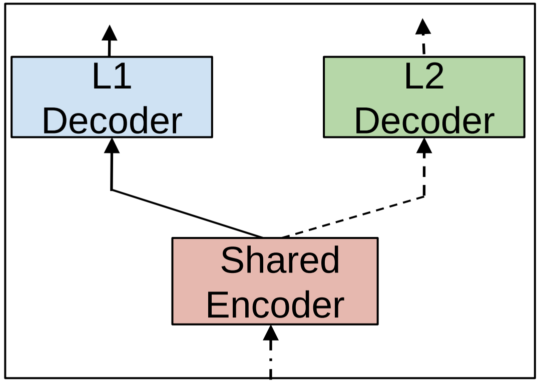
Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

U-NMT: Training with Back-translated data (target to source)

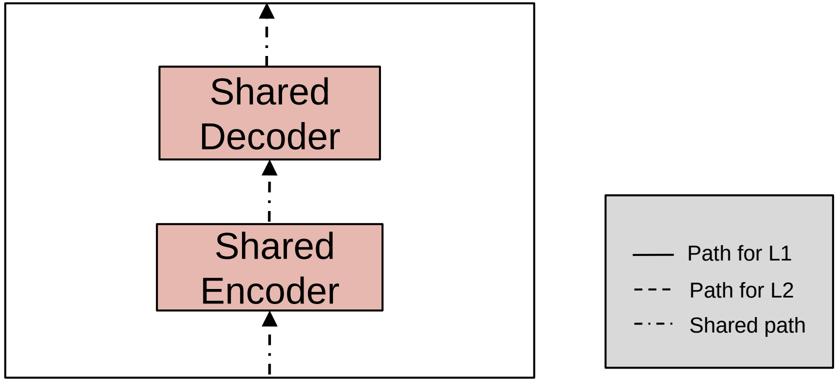


Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

Comparison between two approaches



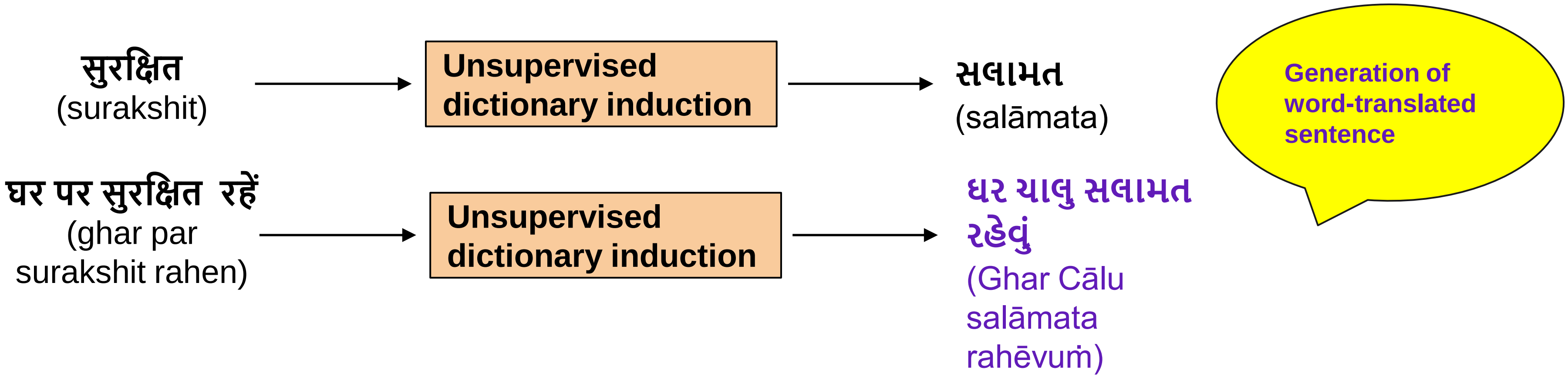
Artexte et al.



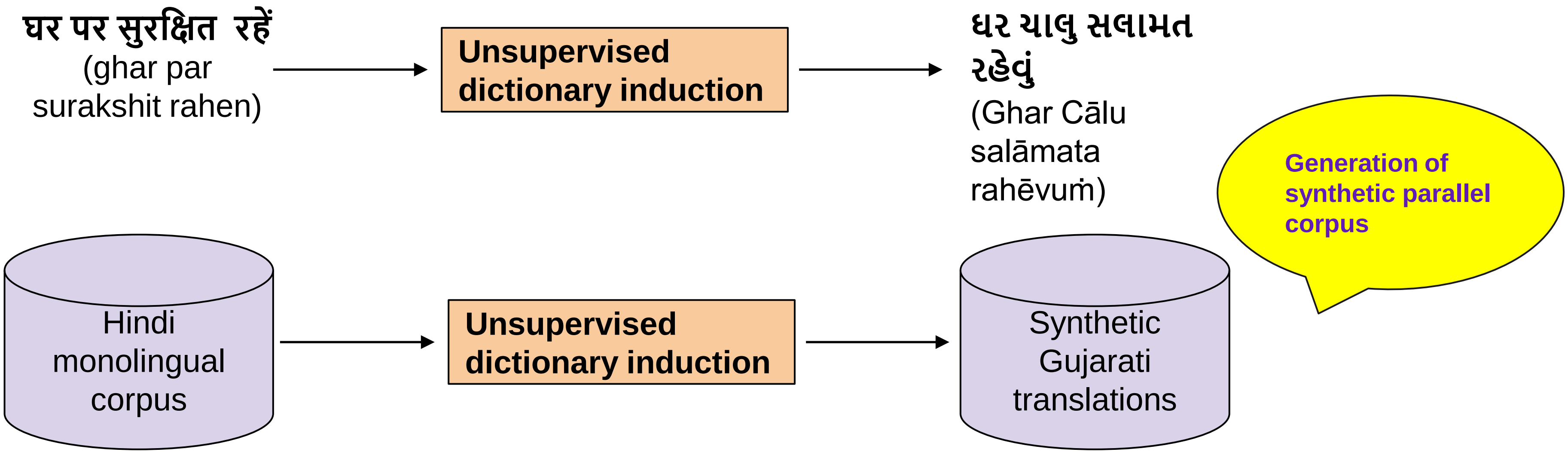
Lample et al.

- Decoders are non-shared for Artexte et al. and shared for Lample et al.
- Lample et al. initialises training with word-by-word translation. [Next few slides]
- Lample et al. uses a language discriminator for encoder representation. It challenges the language invariance nature of encoder representations. [Next subsection]

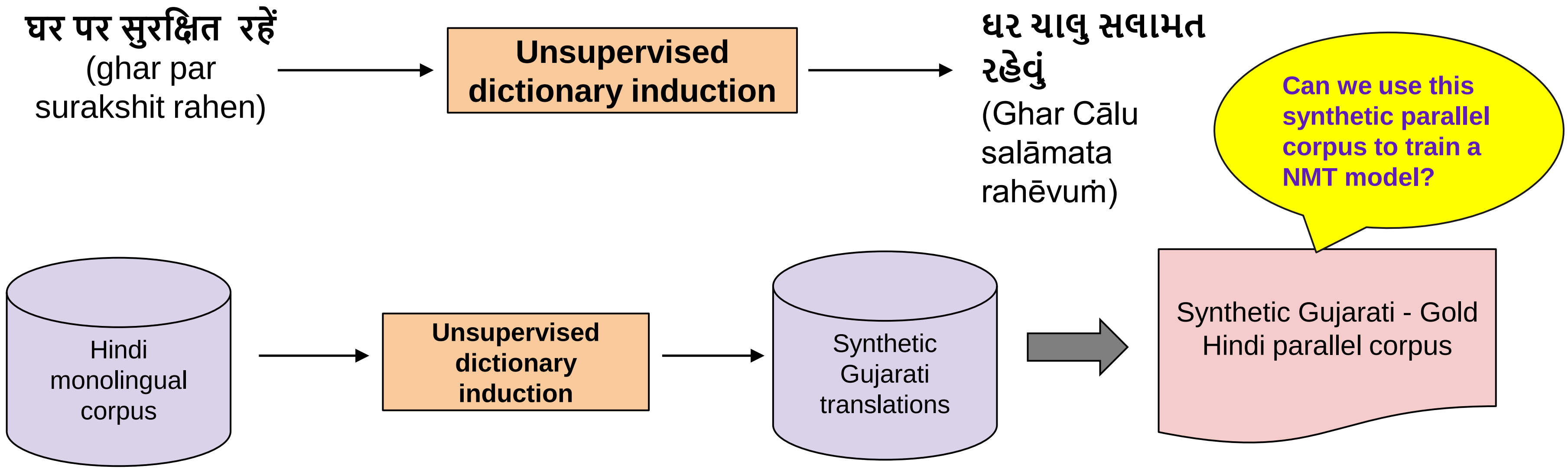
Training with word-by-word translation



Training with word-by-word translation



Training with word-by-word translation



Training with word-by-word translation

घर पर सुरक्षित रहें
(ghar par
surakshit rahen)

→ **Unsupervised dictionary induction** →

ઘર ચાલુ સલામત
રહેવું
(Ghar Cālu
salāmata
rahēvuṁ)

घर पर सुरक्षित रहें
(ghar par
surakshit rahen)

→ **UNMT** →

ઘરે સલામત રહેવું
(Gharē
salāmata
rahēvuṁ)

Generation of sentence translation

Effect of DAE and BT

Author	Approach	Fr → En	En→ Fr	De → En	En → De
Artexte et al. (tested on WMT14)	Emb. nearest neighbour	9.98	6.25	7.07	4.39
	Denoising	7.28	5.33	3.64	2.40
	Denoising + Back-translation	**15.56**	**15.13**	**10.21**	**6.55**
Lample et al. (tested on WMT14 en-fr and WMT16 en-de)	Emb. nearest neighbour	10.09	6.28	10.77	7.06
	Word2word pretraining + Denoising + Back-translation	**15.31**	**15.05**	**13.33**	**9.64**

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).
G. Lample, A. Conneau, L. Denoyer, MA. Ranzato. 2018. Unsupervised Machine Translation With Monolingual Data Only. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

UMT Approaches

1. Unsupervised NMT
2. **GAN for UNMT**
3. Unsupervised SMT
4. Hybrid UMT

Introduction

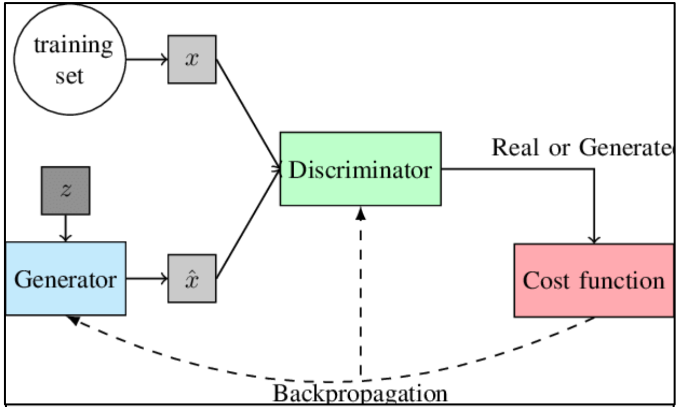
- Use GAN to enhance the language invariance.
- Sharing of the whole model faces difficulty in keeping the diversity of languages.
    - Share module partially

List of papers

1. Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

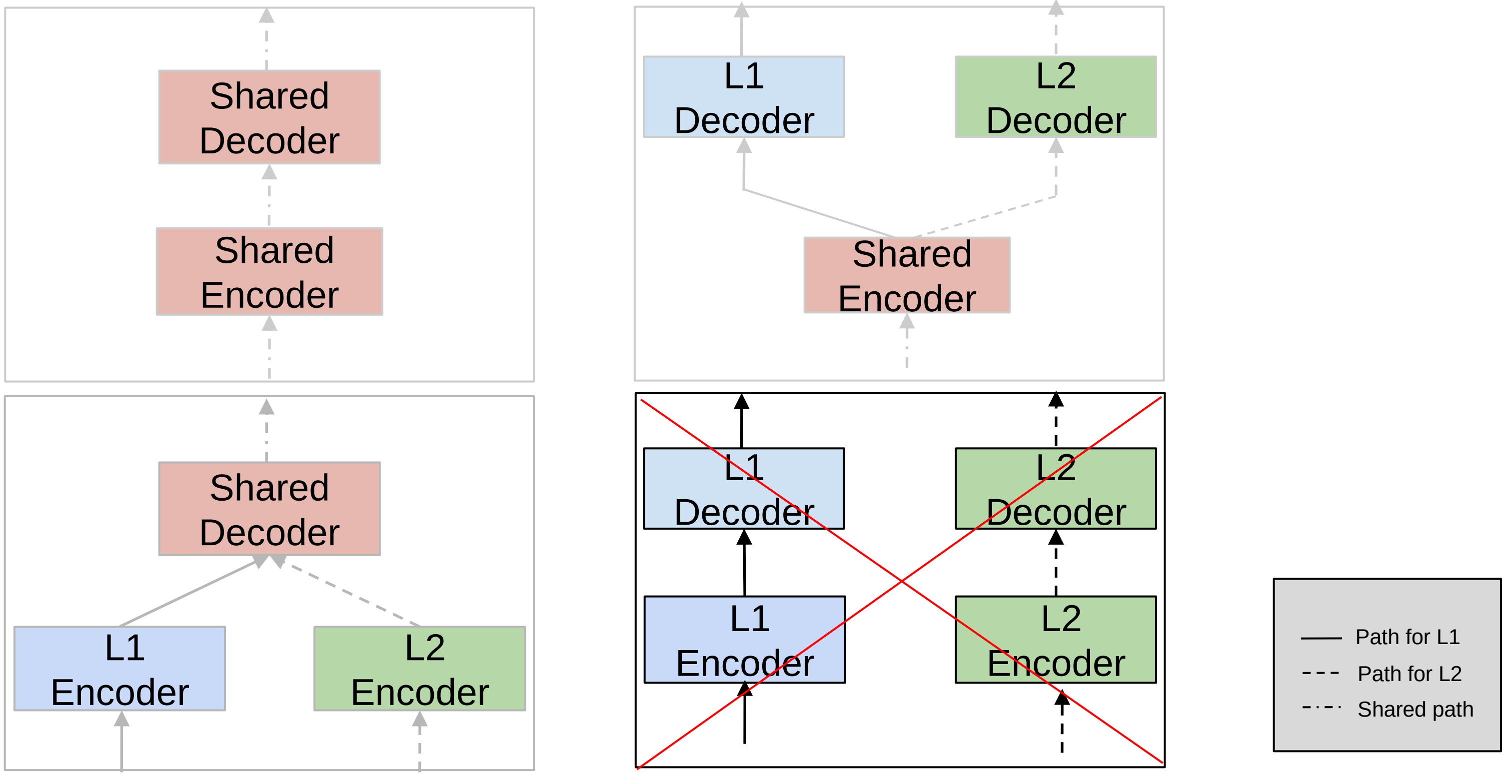
Generative Adversarial Networks (GAN)

- GANs are a clever way of training with two sub-models:
  - Generator model that we train to generate new examples,
  - Discriminator model that tries to classify examples as either real.
- In case of UNMT,
  - Shared encoder is the generator.
  - An extra discriminator module is attached with it to discriminate encoder representations w.r.t. language.

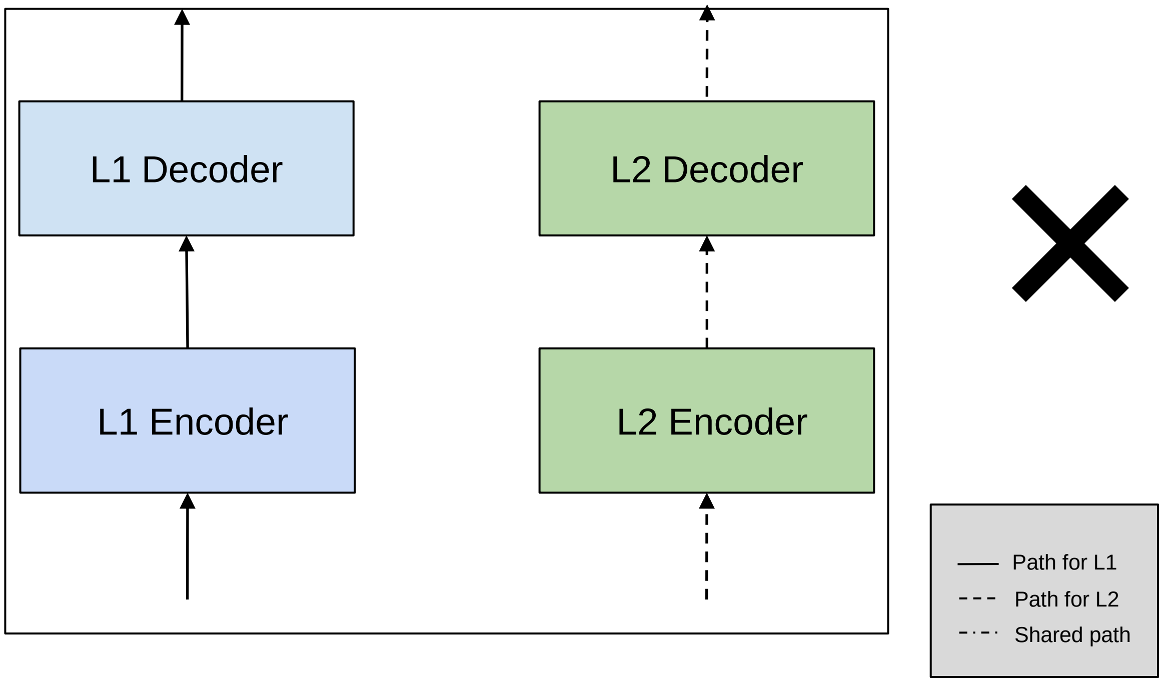


- GAN: Two neural networks (a generative network and a discriminative network) compete with each other to become more accurate in their predictions.

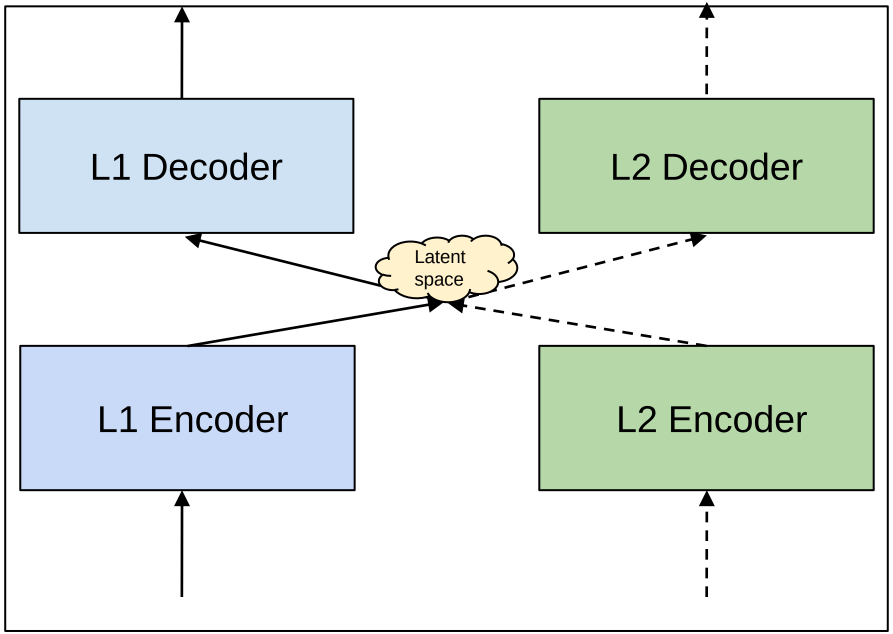
Different parameter sharing strategies

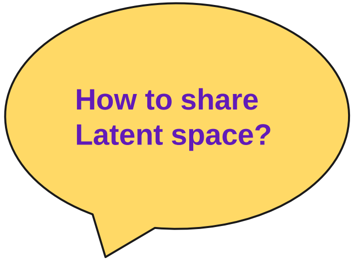


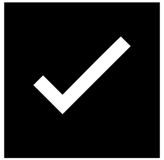
Language specific Encoder-Decoder

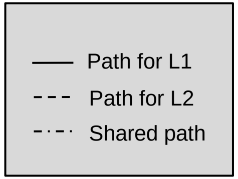


Language specific Encoder-Decoder

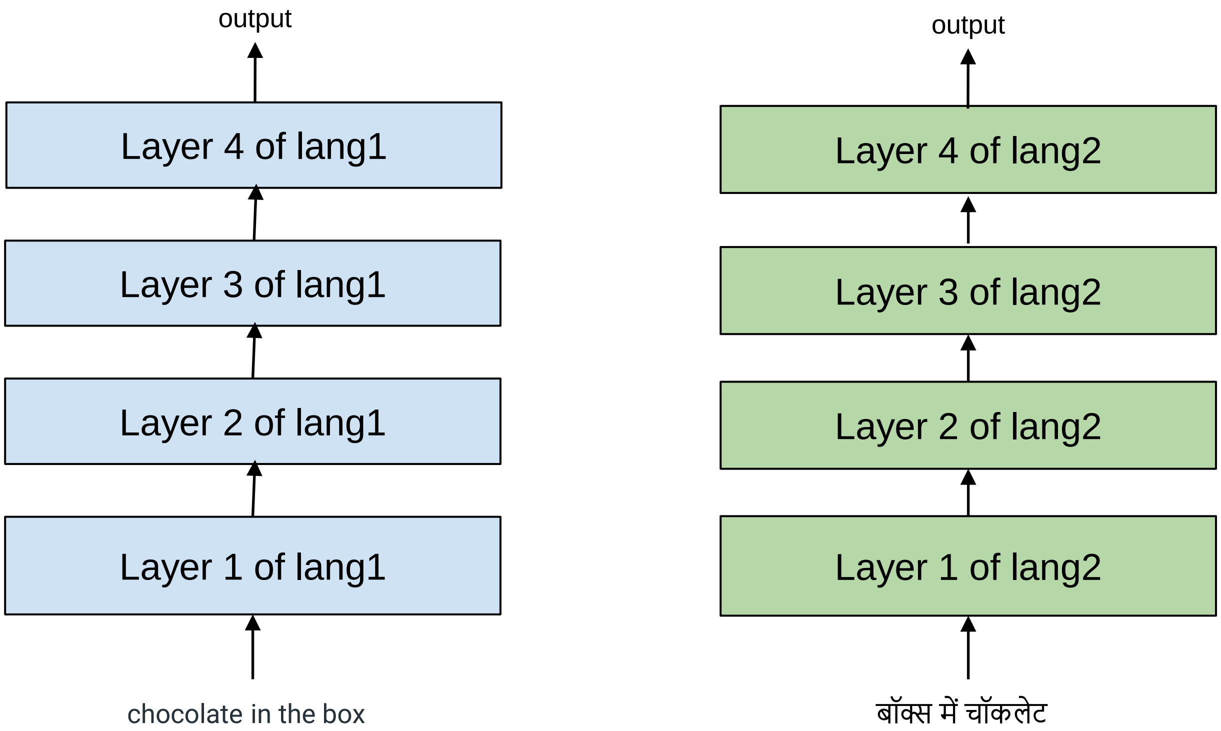




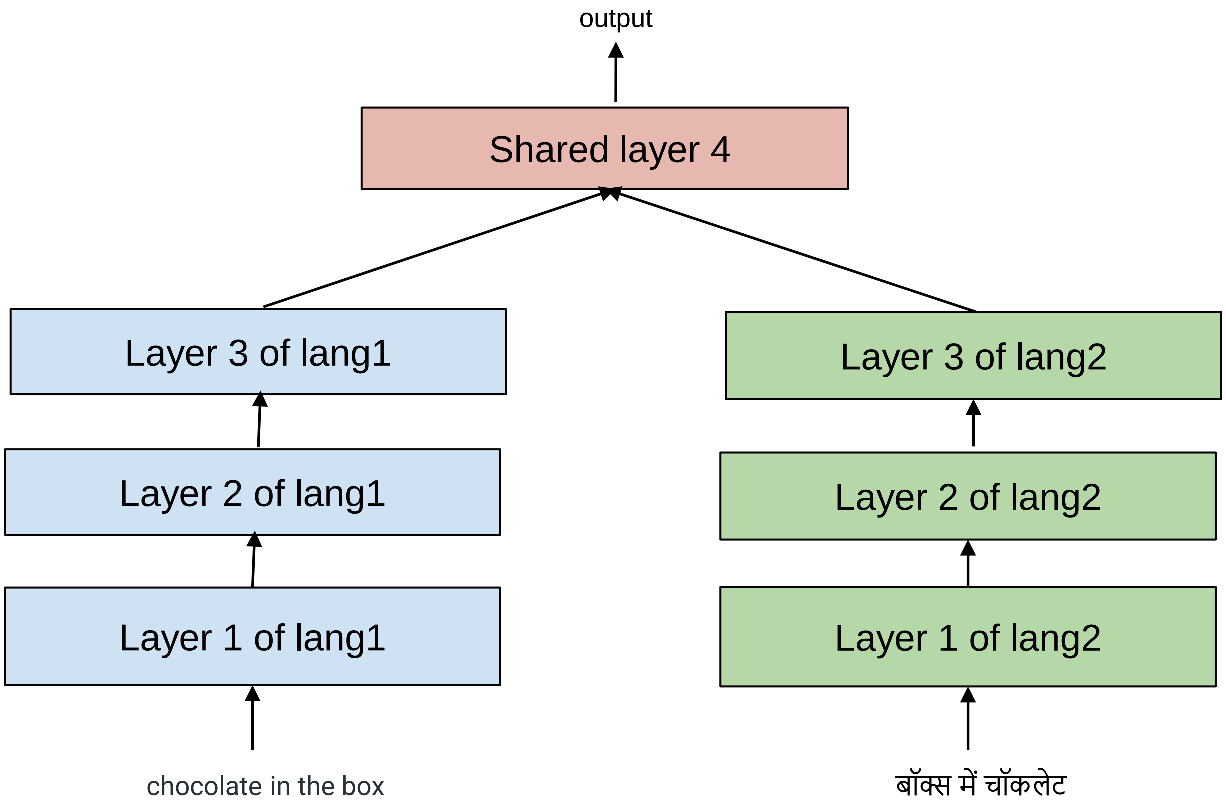




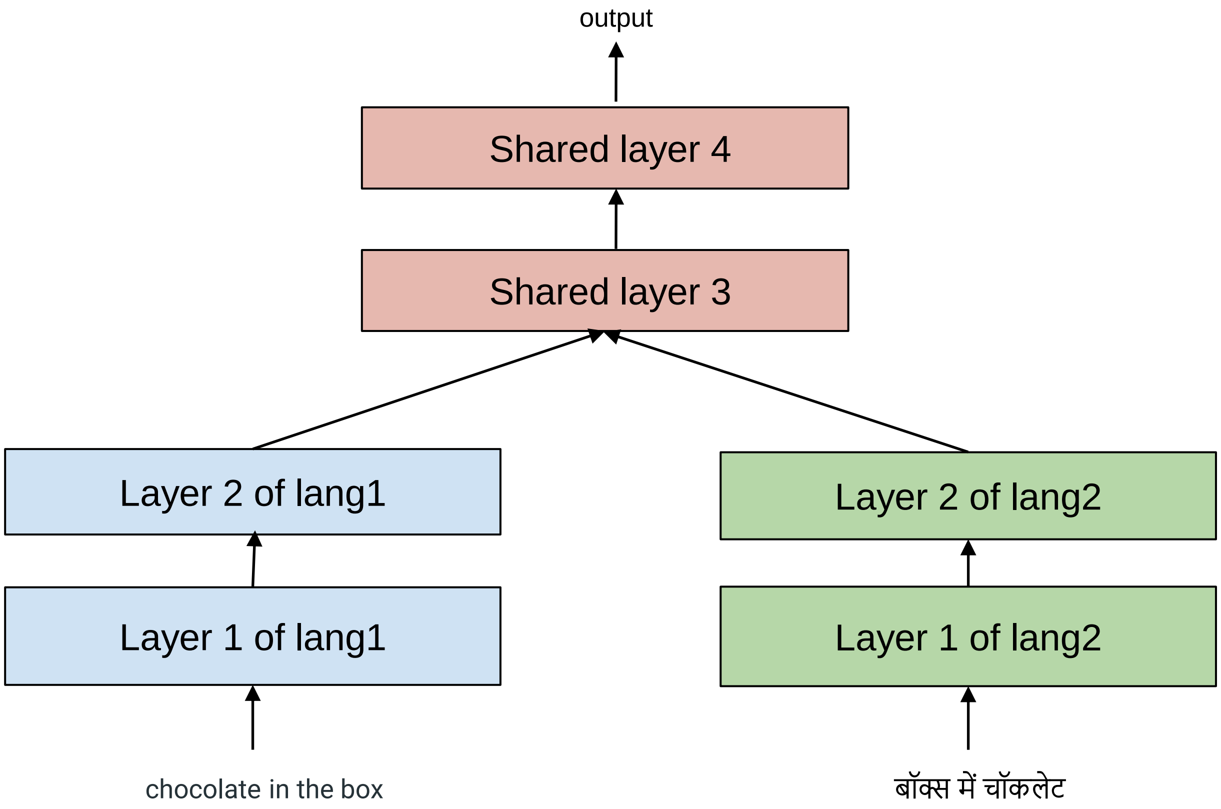
Parameter sharing



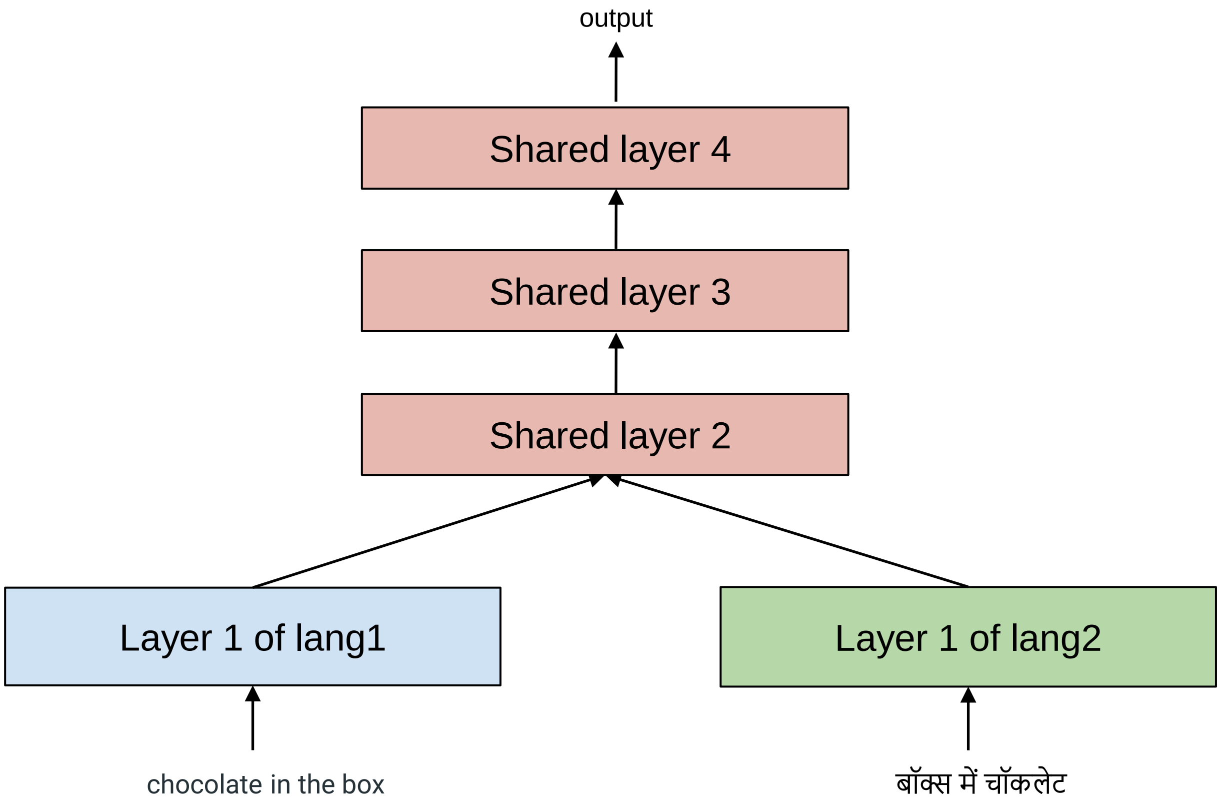
Parameter sharing



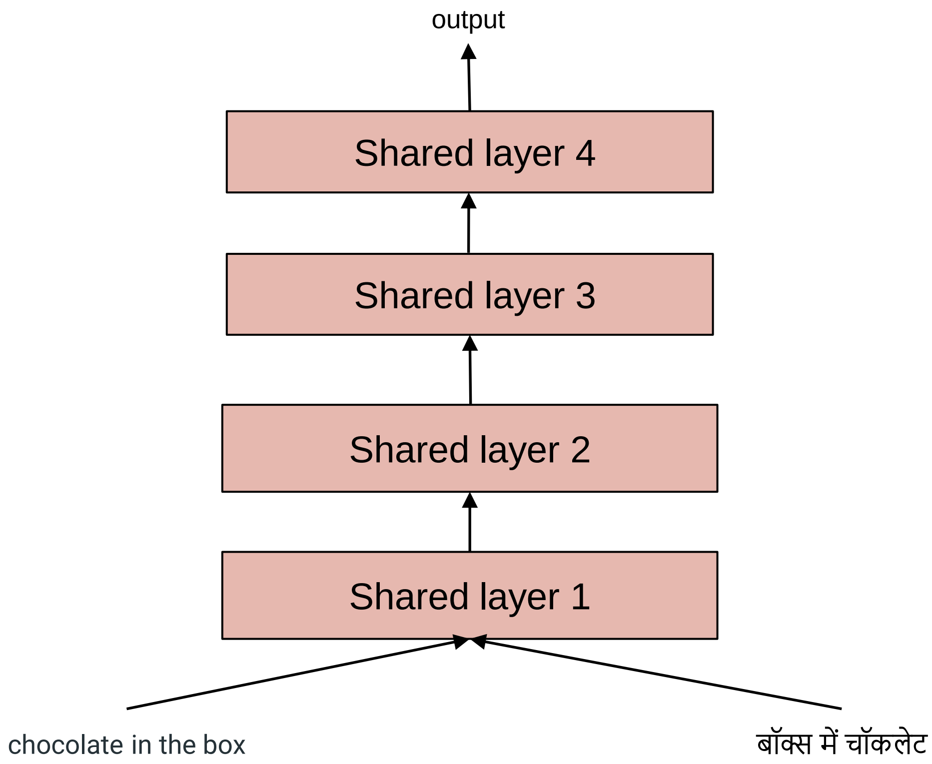
Parameter sharing



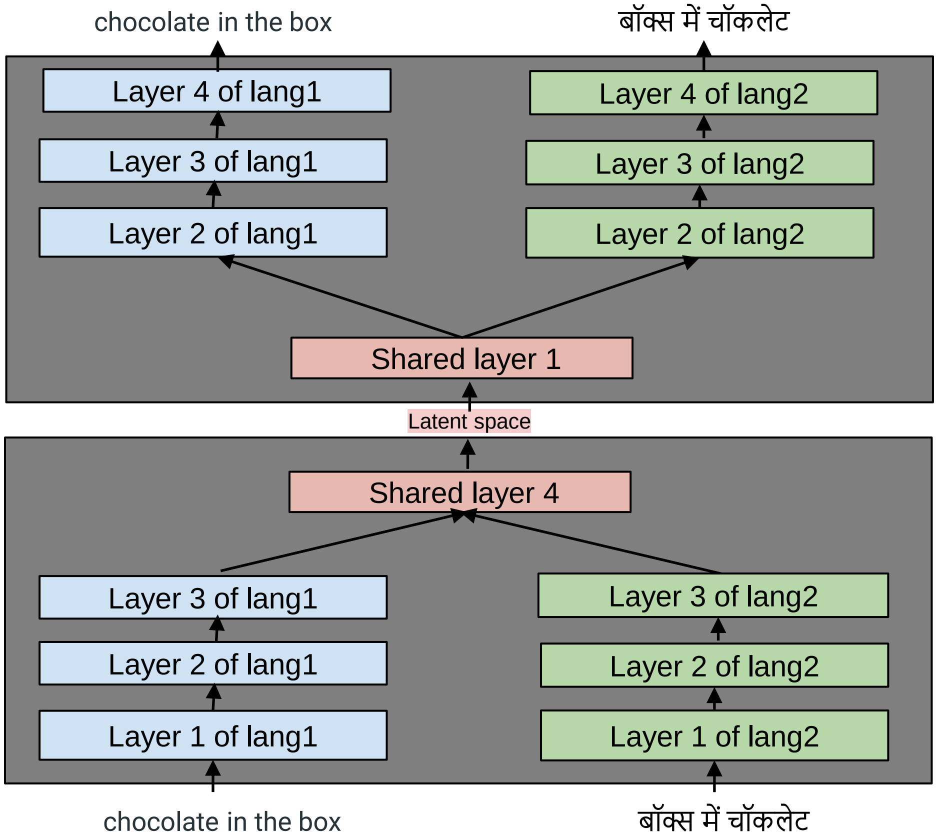
Parameter sharing



Parameter sharing



Architecture with weight-sharing layers



Number of weight-sharing layers vs. BLEU

- In this approach, sharing only 1 layer gives best BLEU scores.
- When sharing is more than 1 layer, the BLEU scores drop.
- This drop is more in case of distant language-pairs when compared to drop in close language-pairs.

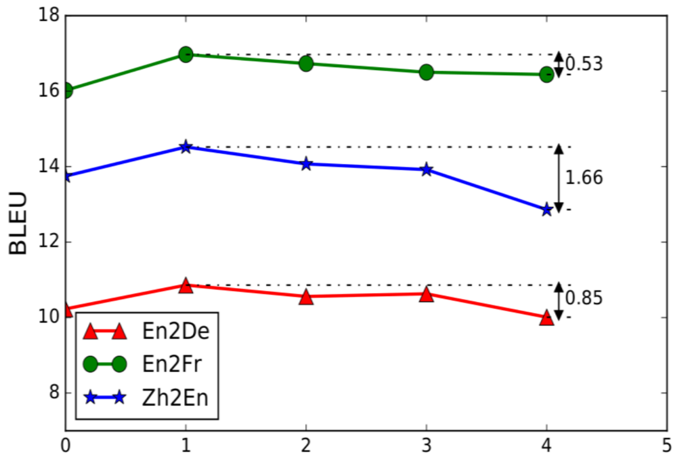
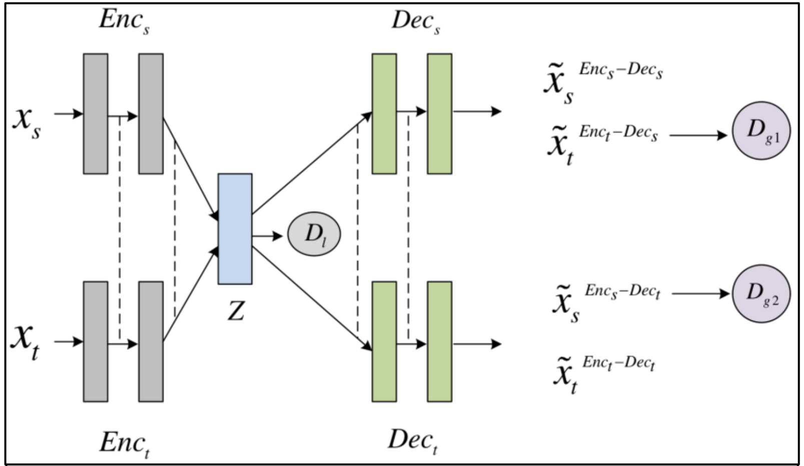


Image source: Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

Weight sharing in UNMT

- When sharing is less, we need GAN to ensure input language invariance of encoder representations and outputs.
- Two types of GAN are used here.
  - Local GAN D_L to ensure input language invariance of encoder representations.
  - Global GAN D_{g1} and D_{g2} to ensure input language invariance of output sentences.



Weight sharing UNMT architecture with GAN

Image source: Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

Results

	en-de	de-en	en-fr	fr-en	zh-en
Supervised	24.07	26.99	30.50	30.21	40.02
Word-by-word	5.85	9.34	3.60	6.80	5.09
Lample et al. (2017)	9.64	13.33	15.05	14.31	-
The proposed approach	**10.86**	**14.62**	**16.97**	**15.58**	**14.52**

Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

UMT Approaches

1. Unsupervised NMT
2. GAN for UNMT
3. **Unsupervised SMT**
4. Hybrid UMT

Introduction

- Components of SMT:
    1) Phrase table
    2) Language model
    3) Reordering model
    4) Word/phrase penalty
    5) Tuning
- Challenges-
    - Phrase table induction without parallel data.
    - Unsupervised Tuning
- Improvement-
    - Iterative refinement
    - Subword information

List of papers

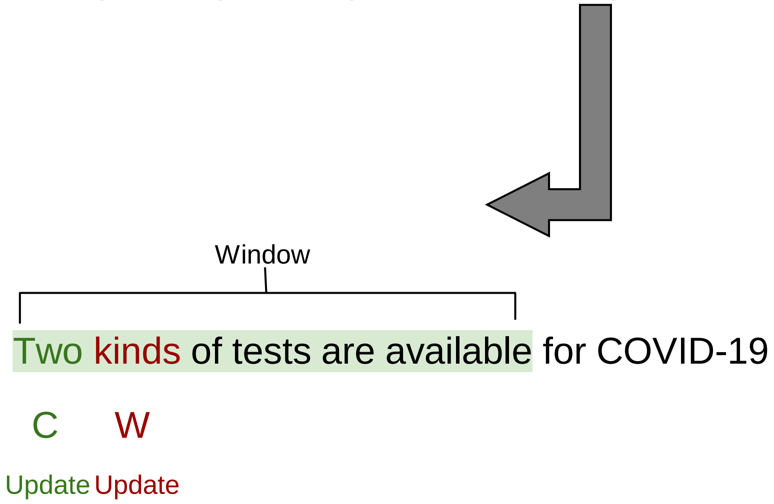
1. Artetxe, M., Labaka, G. and Agirre, E., 2018. Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3632-3642).
2. Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).
3. Artetxe, M., Labaka, G. and Agirre, E., 2019, July. An Effective Approach to Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 194-203).

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

W C

Update Update

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

W C

Update Update

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

W

Update

C

Update

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

W

Update

C

Update

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

P

Update

C

~~Update~~

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

P

Update

C

~~Update~~

Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

Two kinds of tests are available for COVID-19

P

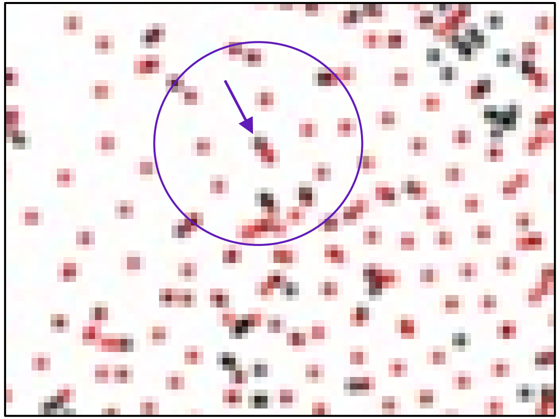
Update

C

~~Update~~

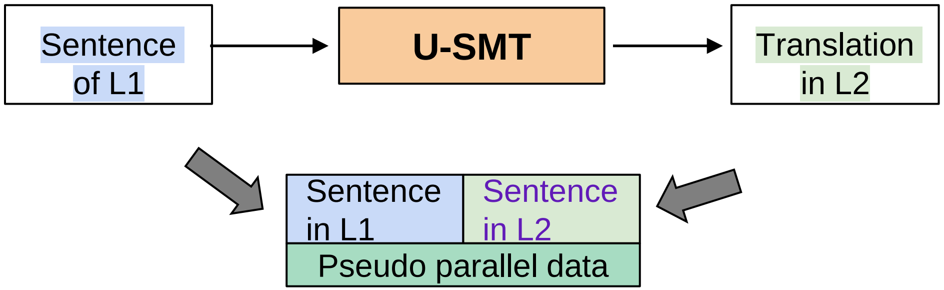
Phrase table induction in an unsupervised way

- Get cross-lingual n-gram embedding.
- Calculate Phrase-translation probabilities.
    - Limit the translation candidates for each source phrase to its 100 nearest neighbors in the target language.
    - Apply the softmax function over the cosine similarities of their respective embeddings.



Unsupervised Tuning

- Tuning with synthetic data.
  - Generate a synthetic parallel corpus.
  - Apply MERT tuning over it iteratively repeating the process in both directions.



- Unsupervised optimization objective:
  - Cyclic loss: The translation of translation of a sentence should be close to the original text.
  - LM loss: We want a fluent sentence in the target language.

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

Iterative refinement

- Generate a synthetic parallel corpus by translating the monolingual corpus with the initial system L1→L2, and train and tune SMT system L2→L1.
  - To accelerate the experiments, use a random subset of 2 million sentences from each monolingual corpus for training.
  - Reuse the original language model, which is trained in the full corpus.
- The process can be repeated iteratively until some convergence criterion is met.

Adding subword information

- We want to favor phrase translation candidates that are similar at the character level.
- Additional weights are added to initial phrase-table.
    - Unlike lexical weightings it use a character-level similarity function instead of word translation probabilities.

$$\boxed{\mathrm{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \mathrm{sim}(\bar{f}_i, \bar{e}_j)\right)}$$

Results

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Unsupervised SMT	21.16	20.13	13.86	10.59	18.01	13.22
+ unsupervised tuning	22.17	22.22	14.73	10.64	18.21	13.12
+ iterative refinement (it1)	24.81	26.53	16.01	13.45	20.76	16.94
+ iterative refinement (it2)	**26.13**	**26.57**	17.30	13.95	22.80	18.18
+ iterative refinement (it3)	25.87	26.22	**17.43**	**14.08**	**23.05**	**18.23**

Artetxe, M., Labaka, G. and Agirre, E., 2018. Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3632-3642).

UMT Approaches

1. Unsupervised NMT
2. GAN for UNMT
3. Unsupervised SMT
4. **Hybrid UMT**

Introduction

- We can combine UNMT and USMT in two ways.
    - USMT followed by UNMT.
    - UNMT followed by USMT.

List of papers

1. Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).

1. Artetxe, M., Labaka, G. and Agirre, E., 2019, July. An Effective Approach to Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 194-203).

USMT followed by UNMT Vs. UNMT followed by USMT

USMT followed by UNMT wins.

- USMT followed by UNMT:
  - Generate pseudo parallel data with USMT.
  - Initialise UNMT system with the pseudo parallel data.
- UNMT followed by USMT:
  - Generate pseudo parallel data with UNMT.
  - Initialise USMT system with the pseudo parallel data.

WMT 14/16	En→Fr	Fr→En	En→De	De→En
NMT + PBSMT	27.1	26.3	17.5	22.1
PBSMT + NMT	**27.6**	**27.7**	**20.2**	**25.2**

Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).

Pre-training approaches for Unsupervised NMT

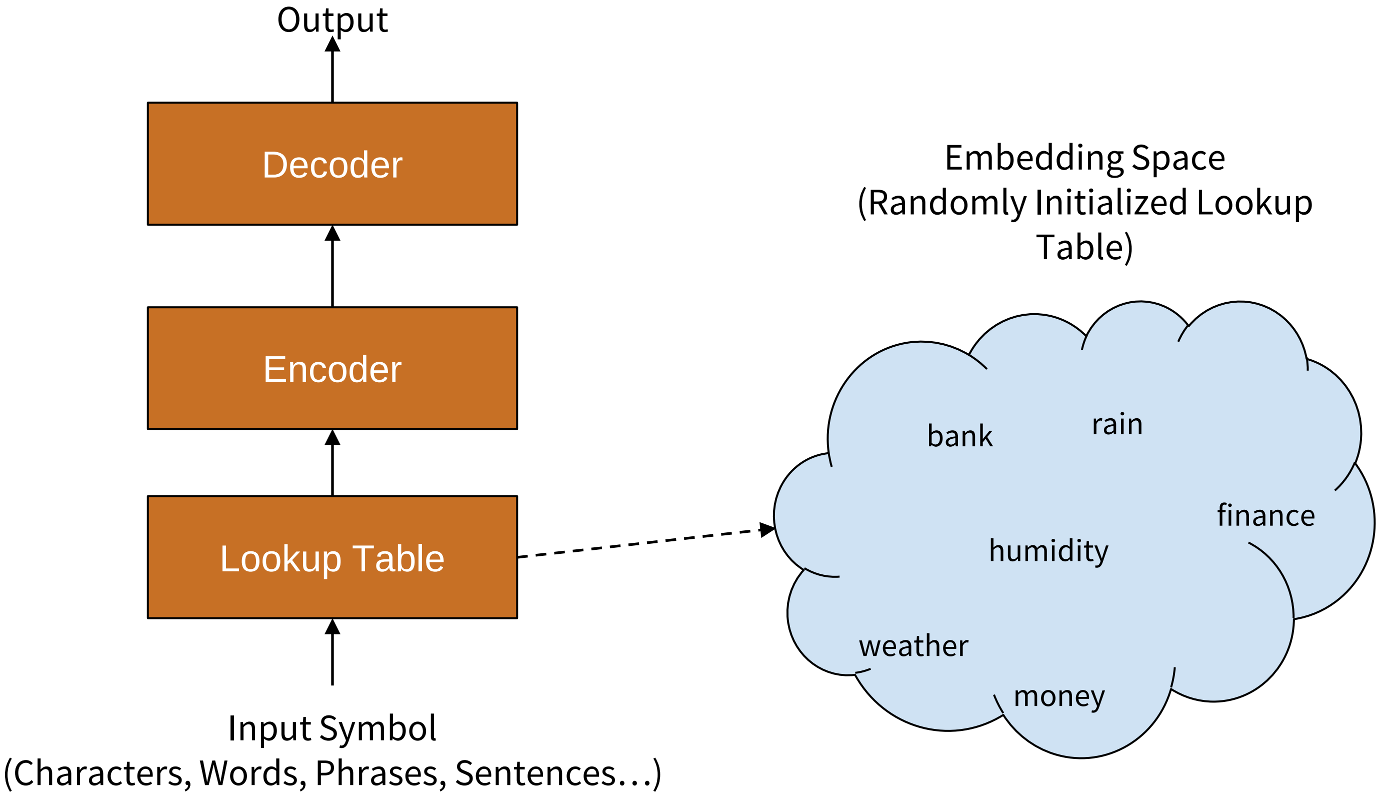
XLM, CMLM, MASS, BART, mBART

XLM

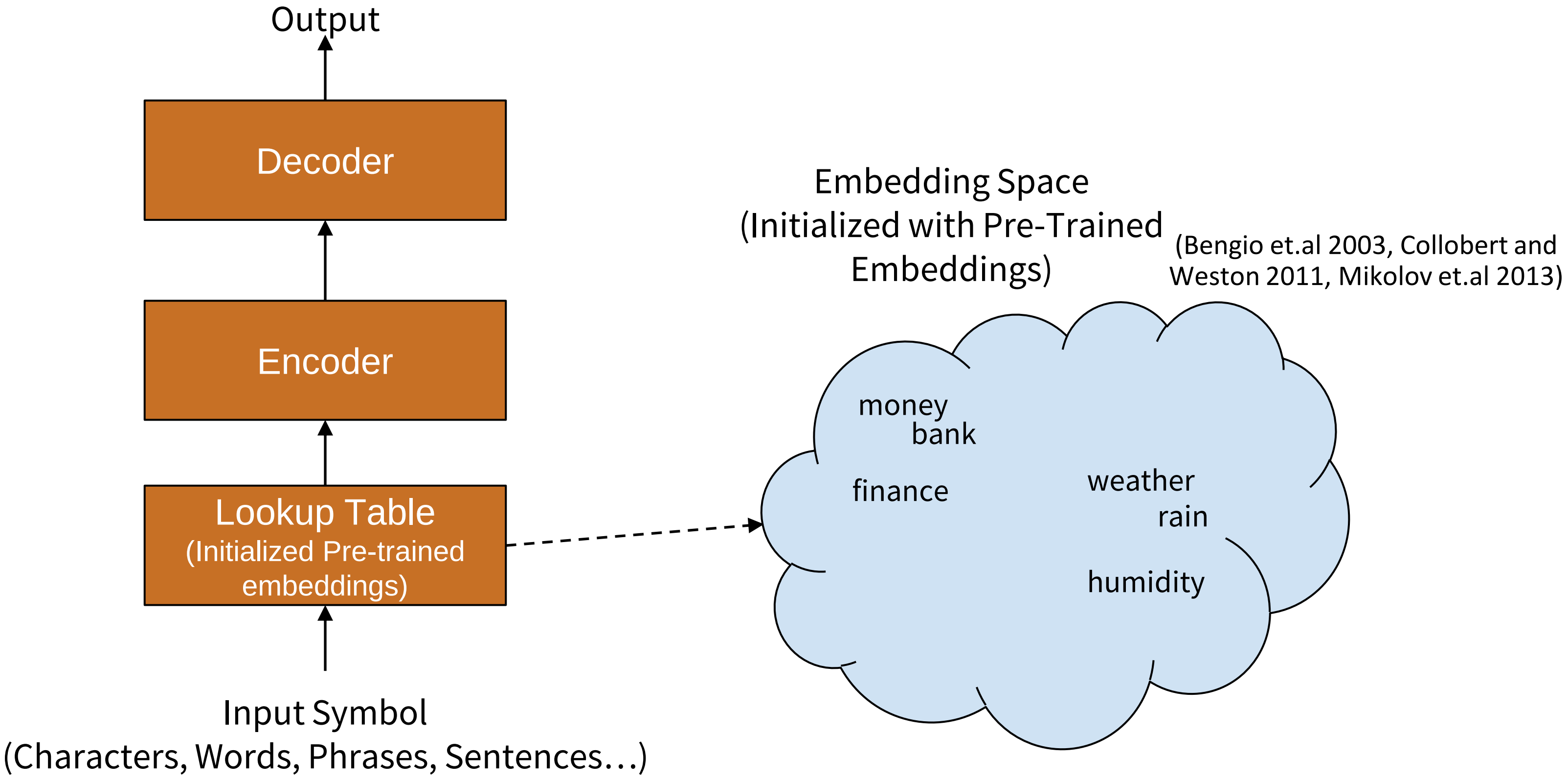
Cross-lingual Language Modelling Pre-Training

Cross-lingual Language Model Pretraining, Advances in Neural Information Processing Systems. 2019.

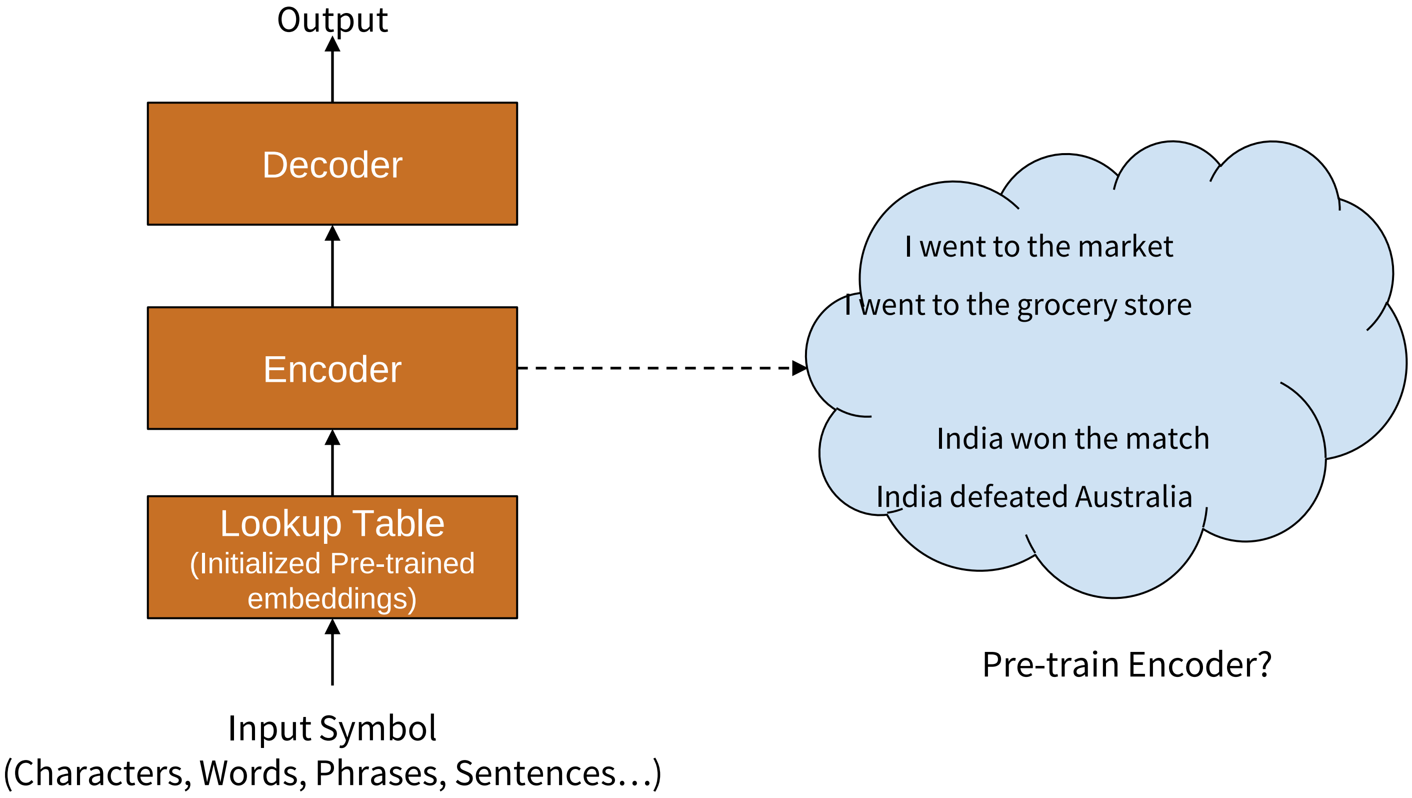
Typical Deep Learning Module



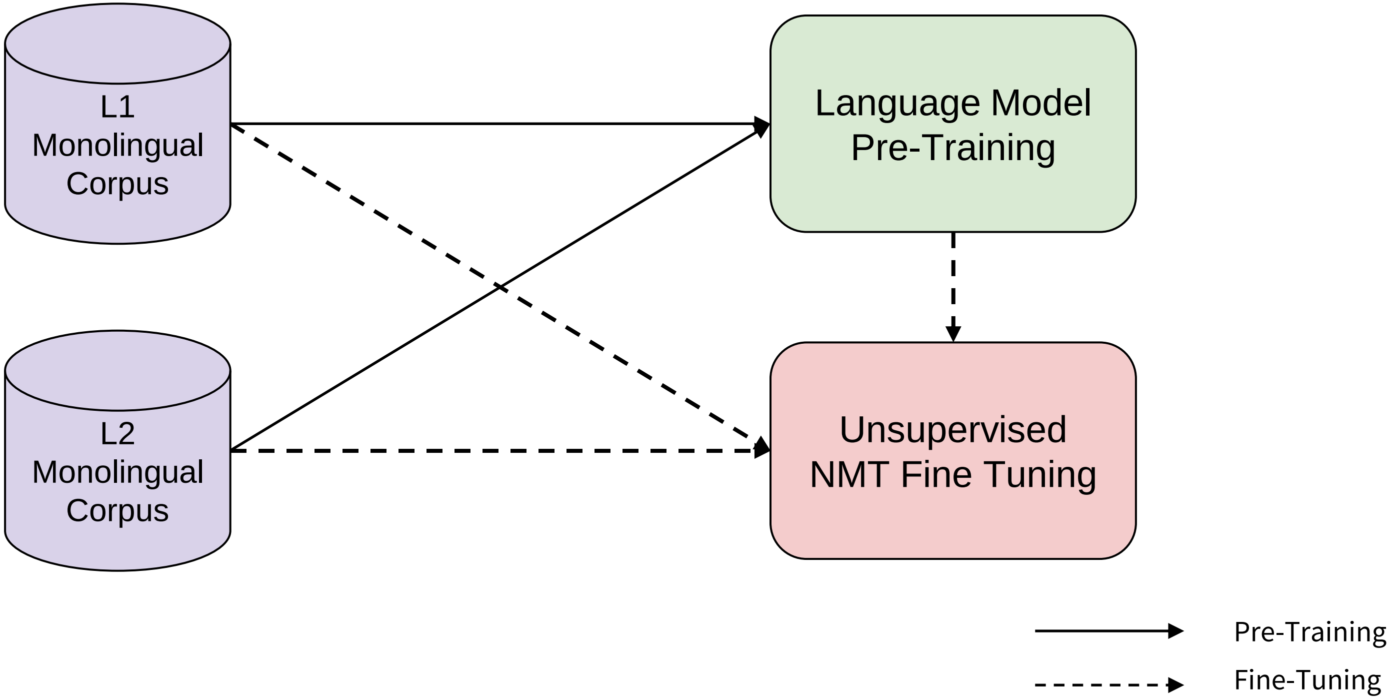
Typical Deep Learning Module



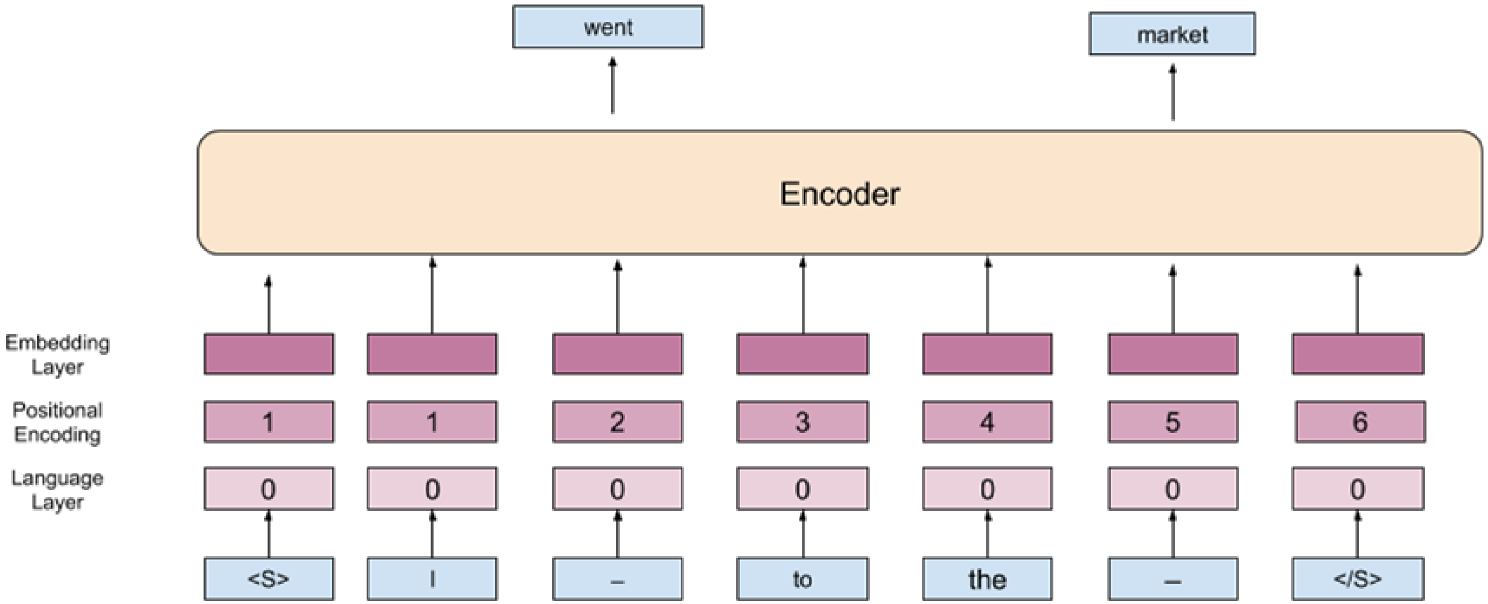
Typical Deep Learning Module



General Framework



XLM Pre-Training



XLM Fine Tuning

- Perform fine-tuning using
  - Iterative back-translation
  - Denoising auto-encoding
- Alternate between the two objective
- Denoising auto-encoding helps in better training of the decoder

XLM: Results

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

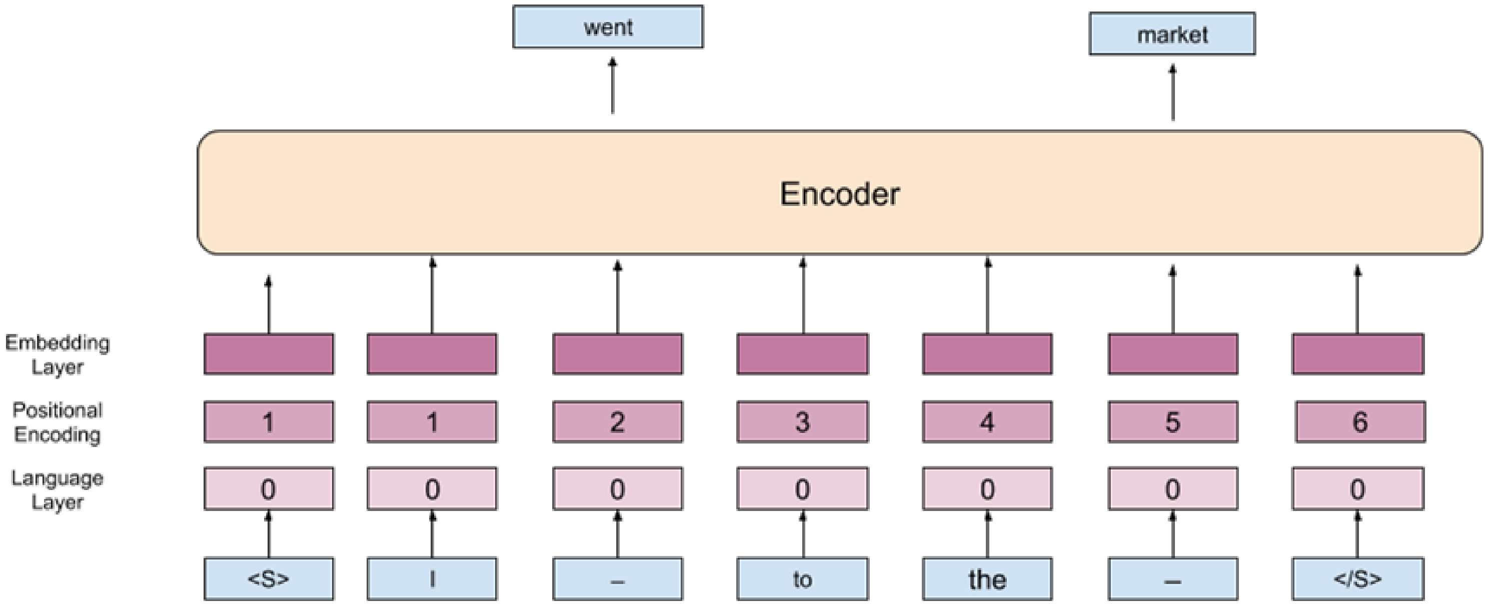
- MLM objective results in better BLEU score compared to Causal Language Modeling (CLM) objective

CMLM

Cross-lingual Masked Language Modelling

Explicit Cross-lingual Pre-training for Unsupervised Machine Translation, EMNLP-IJCNLP 2019

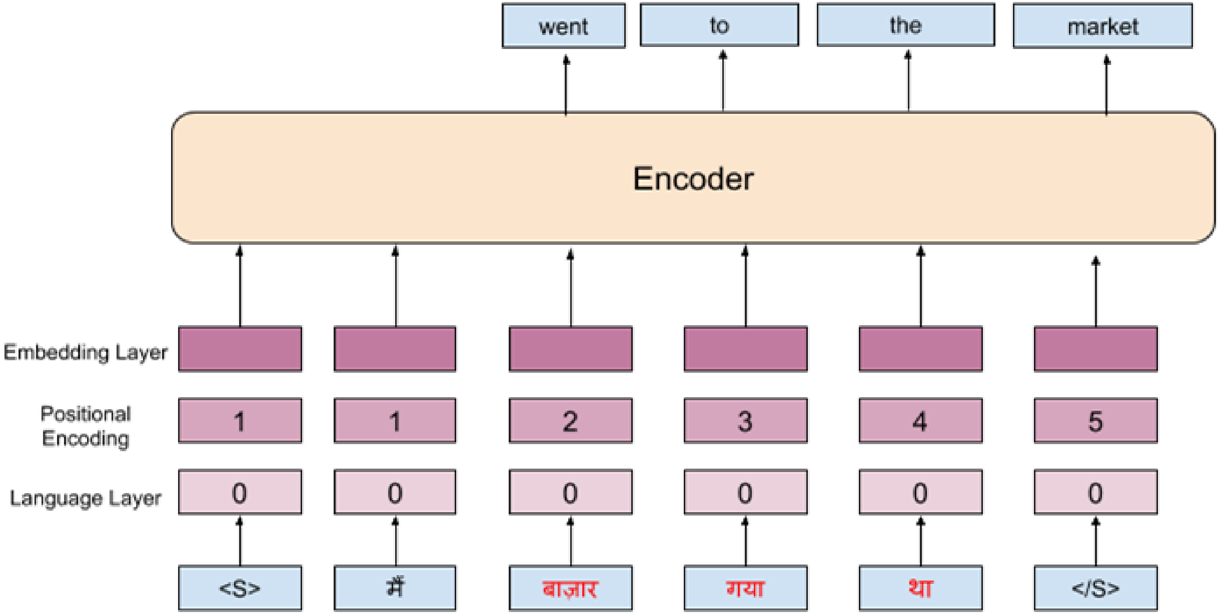
MLM (Devlin et.al 2018)



Limitations

- MLM is trained to predict the missing word in the sentence
- Also, joint training on the combined corpus is not a strong signal to learn good multilingual representations
- Provide explicit cross-lingual signals to the model while pre-training

Cross-lingual Masked Language Modelling



Cross-lingual Masked Language Modelling

- Obtain n-gram phrase translations as discussed earlier
- MLM tries to predict the masked words/tokens
- Modify MLM objective to predict the translation of phrases
- **Mismatch between source and target phrase length**

Cross-lingual Masked Language Modelling

Challenges

- The source and target phrases are of unequal length
- For BERT or XLM, the decoder is a linear classifier.
- Introduce IBM model-2 into the objective

$$P(y_1^m \mid x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)$$

ϵ = probability that the translation of x_1^l consists of m tokens

$a(i, |j, l, m)$ = probability that i^{th} source token is aligned to j^{th} target token

Cross-lingual Masked Language Modelling

Modeling

- Introduce IBM model-2 into the objective

$$P(y_1^m \mid x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)$$

ϵ = probability that the translation of x_1^l consists of m tokens

$a(i, |j, l, m)$ = probability that i^{th} source token is aligned to j^{th} target token

- The loss function becomes

$$L_{cmlm} = -\log(\epsilon) - \sum_{j=1}^{m} \log\left(\sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i) \right)$$

Cross-lingual Masked Language Modelling

Modeling

- The loss function becomes

$$L_{cmlm} = -\log(\epsilon) - \sum_{j=1}^{m} \log\left(\sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i) \right)$$

- The gradient becomes:

$$\nabla L = \sum_{j=1}^{m} \frac{a(i \mid j, l, m)\, P(y_j \mid x_i)}{\sum_{i=0}^{l} a(i \mid j, l, m)\, P(y_j \mid x_i)} \nabla \log P(y_j \mid x_i)$$

Cross-lingual Masked Language Modelling

Modeling

- The gradient becomes:

$$\nabla L = \sum_{j=1}^{m} \frac{a(i \mid j, l, m)\, P(y_j \mid x_i)}{\sum_{i=0}^{l} a(i \mid j, l, m)\, P(y_j \mid x_i)} \nabla \, log \, P(y_j \mid x_i)$$

  - a(i, |j, l, m) are approximated using cross-lingual BPE embedding
  - $P(y_j \mid x_i)$ is calculated by passing x_i contextual embedding representation through a linear layer followed by soft-max

Cross-lingual Masked Language Modelling

Algorithm

- Alternate between CMLM and MLM objective
- In MLM objective,
  - 50% of the time randomly choose some source ngrams and replace it with the corresponding translation candidate (pseudo code-switching)
- In CMLM objective,
  - Randomly select 15% of the BPE ngram tokens and replace them by [MASK] 70% of the time
  - Trained to predict the translation candidate in the other language

Cross-lingual Masked Language Modelling

Results

Method	fr2en	en2fr	de2en	en2de	ro2en	en2ro
(Artetxe et al., 2017)	15.6	15.1	-	-	-	-
(Lample et al., 2017)	14.3	15.1	13.3	9.6	-	-
(Artetxe et al., 2018b)	25.9	26.2	23.1	18.2	-	-
(Lample et al., 2018)	27.7	28.1	25.2	20.2	23.9	25.1
(Ren et al., 2019)	28.9	29.5	26.3	21.7	-	-
(Lample and Conneau, 2019)	33.3	33.4	34.3	26.4	31.8	33.3
Iter 1 (CMLM)	34.8	34.9	35.5	**27.9**	33.6	34.7
Iter 2	**34.9**	**35.4**	**35.6**	27.7	**34.1**	**34.9**

CMLM

Cross-lingual Masked Language Modelling

Ablation Study

CMLM: Ablation Study

- Role of n-gram masking
- Influence of translation prediction

	fr2en	en2fr	de2en	en2de
CMLM + MLM	**34.8**	**34.9**	**35.5**	**27.9**
CMLM	34.1	34.3	35.1	27.2
- translation prediction	33.7	33.9	34.8	26.6
- - n-gram mask	33.3	33.4	34.3	26.4

CMLM + MLM means we use L_{pre} as the pre-training loss;
CMLM means we only use L_{cmlm} as the pre-training loss;
-- translation prediction predict the masked n-grams rather than their translation candidates;
- - n-gram mask randomly mask BPE tokens rather than n-grams based on **-- translation prediction** during pre-training, which degrades our method to XLM.

MASS

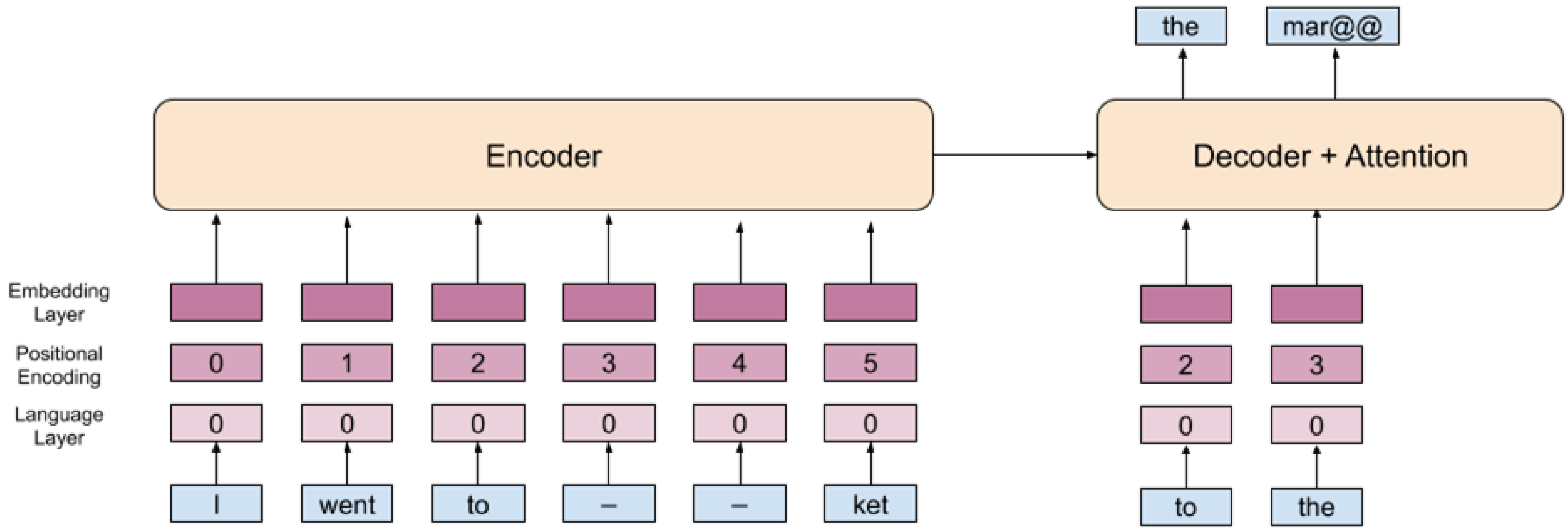
Masked Sequence to Sequence pretraining

MASS: Masked Sequence to Sequence Pre-training for Language Generation, ICML, Song et.al 2019

MASS (Song et.al 2019)

- XLM objective predicts the masked word in the sentence
- However, for U-NMT we need to generate a sequence
- This disconnect between pre-training and fine-tuning objective could limit the potential of unsupervised pre-training
- MASS extends XLM objective to include text segments
- Given a sentence, randomly mask k% of the text segment
- The decoder has to generate the masked text segment now

MASS Pre-Training



MASS Fine-Tuning

- Perform fine-tuning using iterative back-translation
- Unlike XLM which had
  - iterative back-translation
  - Denoising auto-encoding

MASS (Song et.al 2019)

Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	33.30	31.80
MASS	6-layer Transformer	**37.50**	**34.90**	**28.30**	**35.20**	**35.20**	**33.10**

Table 2. The BLEU score comparisons between MASS and the previous works on unsupervised NMT. Results on en-fr and fr-en pairs are reported on *newstest2014* and the others are on *newstest2016*. Since XLM uses different combinations of MLM and CLM in the encoder and decoder, we report the highest BLEU score for XLM on each language pair.

MASS

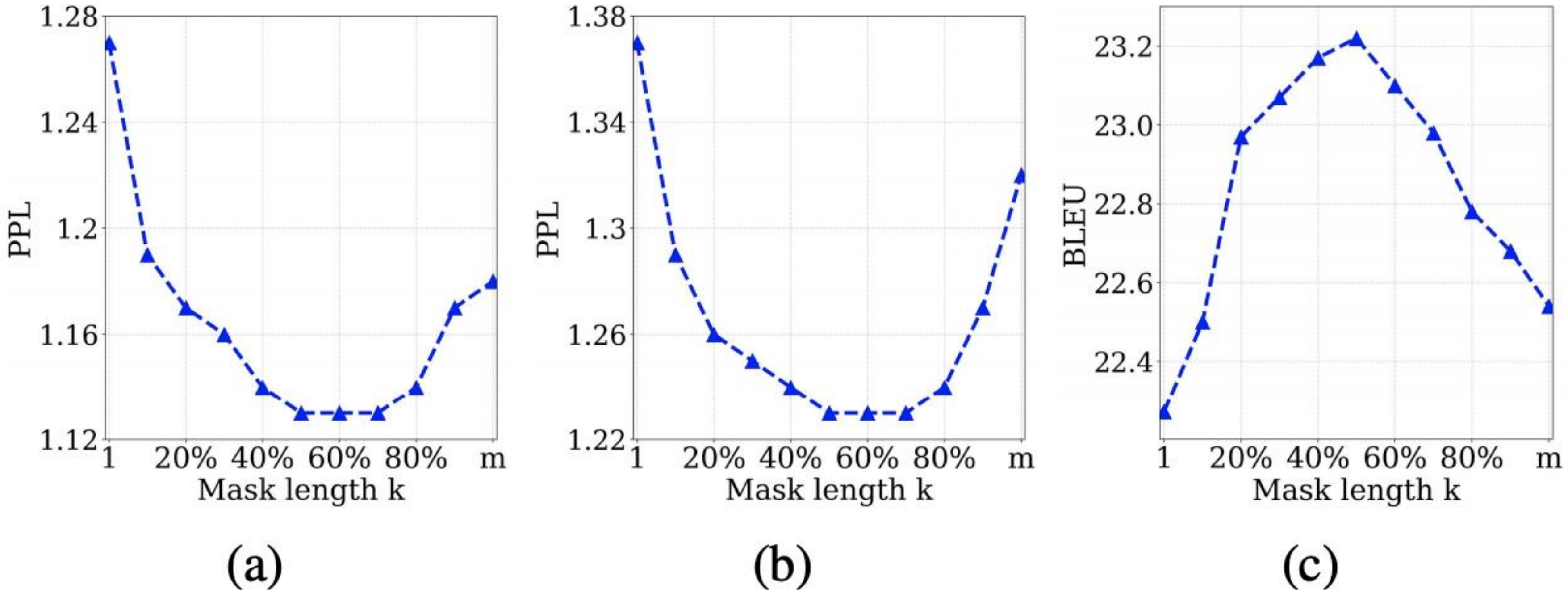
Masked Sequence to Sequence pretraining

Role of hyper-parameters

MASS Hyper-parameters

- Percentage of ngram tokens in a sentence to be masked (**masking length**)
    - Consider the input sentence, **X** = ***I went to the market yesterday night***
    - Let ***to the market yesterday*** be the text segment selected for masking
    - Default value is 50% of the input sentence
    - However, not all tokens ***to the market yesterday*** are masked
- Given a text fragment $x_i, \ldots, x_j$ of length m selected for masking (Word selection)
    - k% of the tokens are selected for masking (**mask probability**)
    - l% of the tokens are replaced by random tokens (**replace probability**)
    - (100 - (k + l)) of the tokens are retained (**keep probability**)
    - Default values are k = 80%, l = 10%

MASS (Song et.al 2019): Role of Masking Length



The performances of MASS with different masked lengths k, in both pre-training and fine-tuning stages, which include: the PPL of the pre-trained model on English (Figure a) and French (Figure b) sentences from WMT newstest2013 on English-French translation; the BLEU score of unsupervised English-French translation on WMT newstest2013 (Figure c)

MASS Hyper-parameters

मै तो आपके घर से चाय का पत्ती मांगने आयी हूँ

mai to Apake ghara se chAya kA pattI mAMgane AyI hU.N

Select randomly 50% of the consecutive tokens for masking

मै तो आपके _ _ _ _ _ _ मांगने आयी हूँ

80% of the selected tokens are **masked**, 10% **randomly replaced**

मै तो आपके _ से _ पीना _ मांगने आयी हूँ

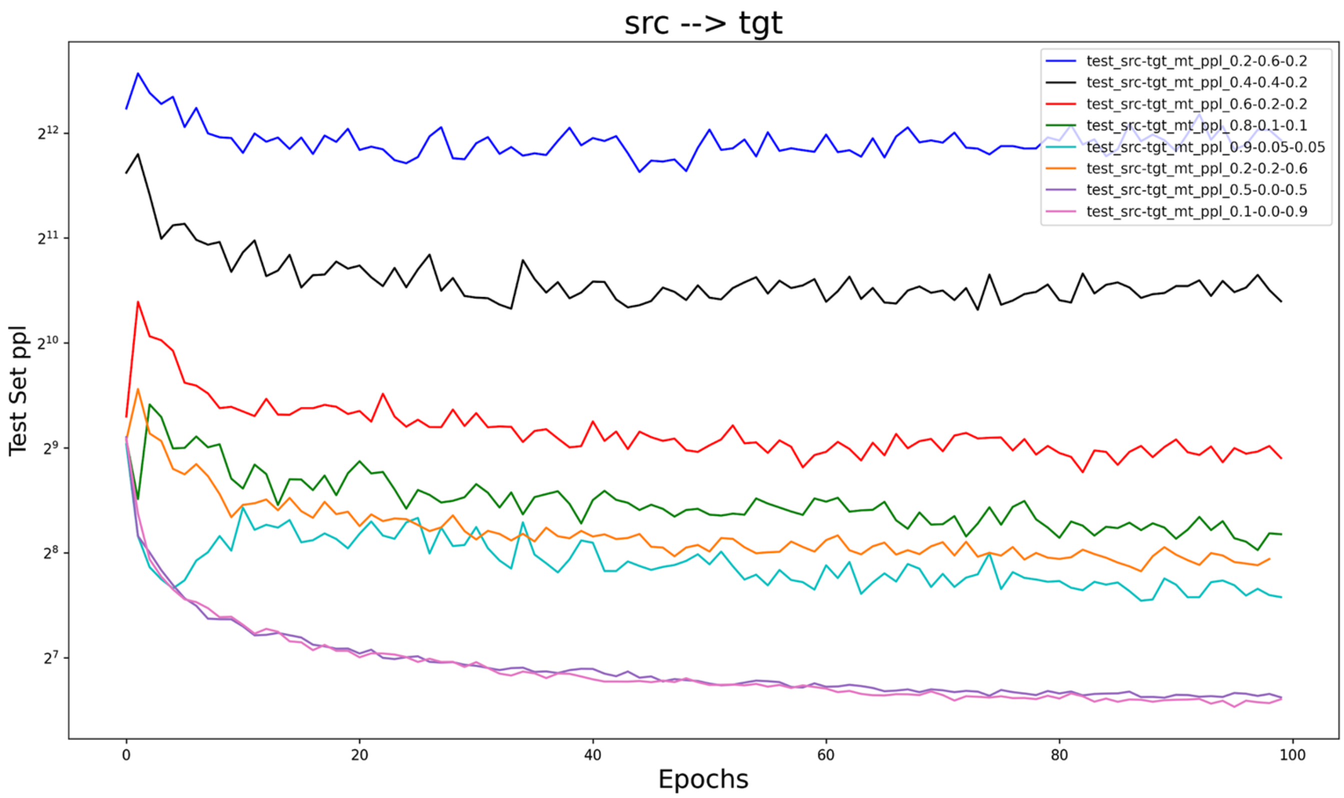
घर से चाय का पत्ती

Output to be generated

MASS: Word Selection Hyper-parameters

Configuration	%age Masked	%age Retained	%age Randomly replaced
1	20	60	20
2	40	40	20
3	60	20	20
4	80	10	10
5	90	5	5
6	20	20	60
7	50	-	50
8	10	-	90

MASS (Song et.al 2019): Word Selection Hyper-parameters



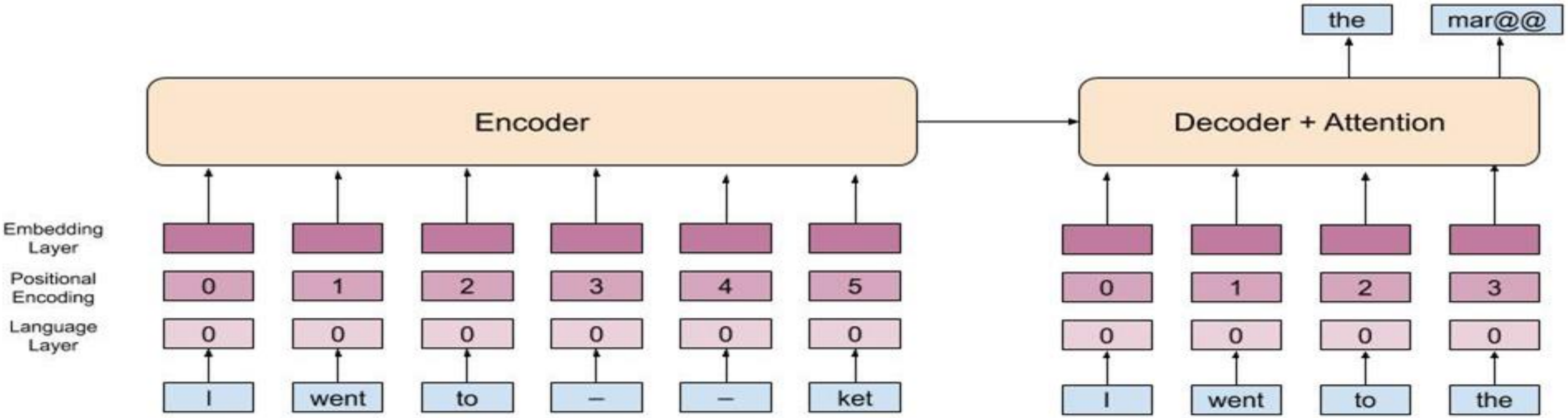
MASS: Word Selection Hyper-parameters

Configuration	%age Masked	%age Retained	%age Randomly replaced	Comments
1	20	60	20	Auto-encoder
2	40	40	20	Auto-encoder
3	60	20	20	Auto-encoder
4	80	10	10	Recommended
5	90	5	5	Recommended
6	20	20	60	Unable to generate translations. But perplexity is low (Better for other tasks?)
7	50	-	50	
8	10	-	90	

MASS (Song et.al 2019): Role of Masking Tokens

- Consider the input sentence, **X** = ***I went to the market yesterday night***
- Let ***to the market yesterday*** be the text segment selected for masking
- The input to the encoder is ***I went _ _ _ _ night***
- The input to the decoder (previous token) is ***went to the market***
  - Why mask consecutive tokens and not discrete tokens? (***Discrete***)
  - Why not feed all the input tokens to the decoder (similar to previous target word in NMT)? (***feed***)

Feeding Input Tokens



MASS (Song et.al 2019): Role of Masking Tokens

Method	BLEU	Method	BLEU	Method	BLEU
Discrete	36.9	*Feed*	35.3	MASS	37.5

The comparison between MASS and the ablation methods in terms of BLEU score on the unsupervised en-fr translation

BART and mBART

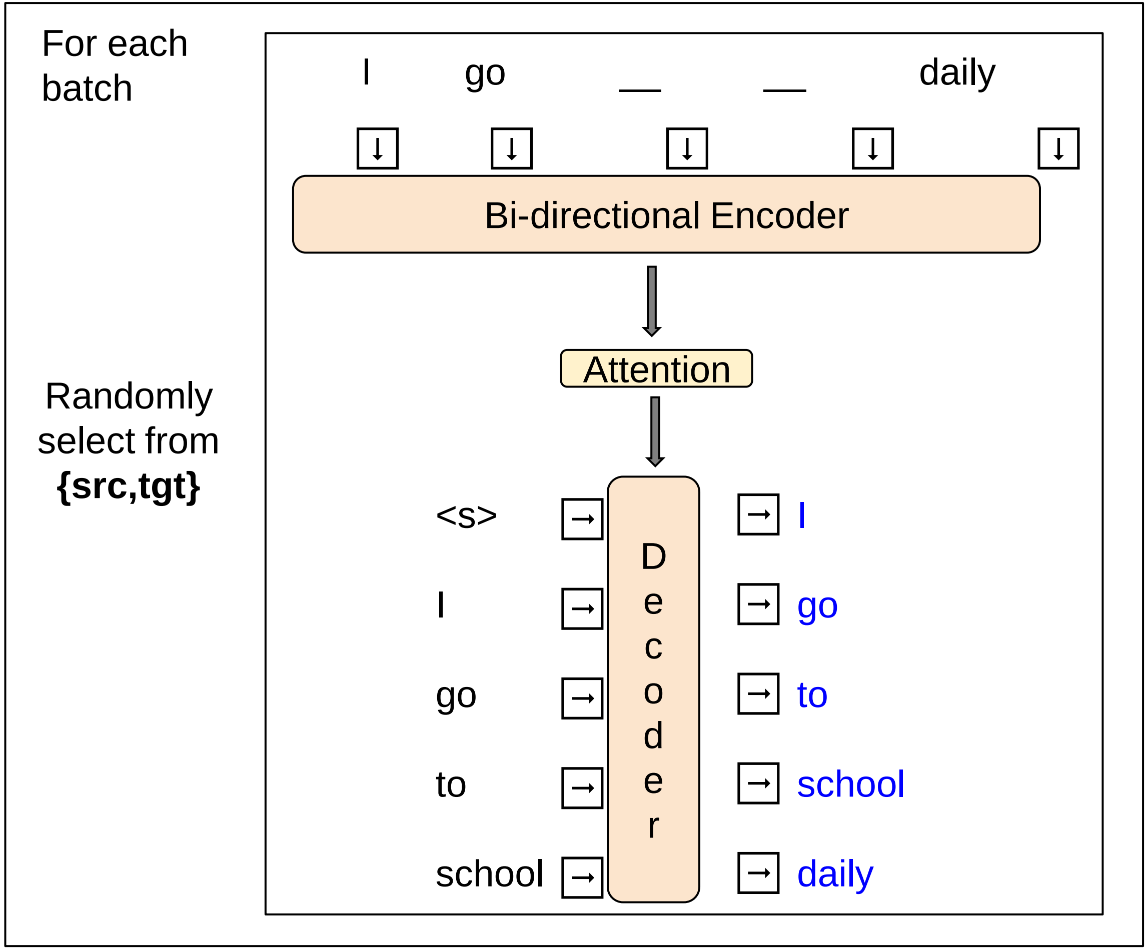
BART: Denoising Sequence to Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, ACL 2020, (Lewis et al 2020)

Multilingual denoising pre-training for Neural Machine Translation, 2020, (Liu et al 2020)

BART Pretraining

- Trained by
  - Corrupting text with an arbitrary noising function
  - Learning a model to reconstruct the original text.
- Denoising full text
- Multi-sentence level

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL 2020)*



BART pretraining (possible noising steps) (Lewis et al. 2020)

My name is John. I go to school daily.

Original document

Token Masking	My _ is John. I __ school daily.
Token deletion	My name John. I go to daily.
Text infilling	My _ John. I go _.
Sentence permutation	I go to school daily. My name is John
Document rotation	name is John. I go to school daily. my

BART noising steps (Lewis et al. 2020)

- Experimented with different noise functions for various tasks
  - Text infilling + Sentence permutation performed the best
    - Remove spans of text and replace with mask tokens
    - Mask 30% of the words in each instance by randomly sampling a span length
    - Permute the order of sentences

mBART (Liu et al 2020)

- A sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in **many languages** using the BART objective
- Unsupervised NMT
  - BART pretraining using monolingual corpora of multiple languages + Iterative Back-Translation

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv2020.

mBART (Liu et al 2020)

- Pre-training using BART objective on multiple languages

Model	Similar Pairs				Dissimilar Pairs			
	En-De		En-Ro		En-Ne		En-Si	
	←	→	←	→	←	→	←	→
Random	21.0	17.2	19.4	21.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	31.8	33.3	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	**33.1**	**35.2**	-	-	-	-
mBART	34.0	**29.8**	30.5	35.0	**10.0**	**4.4**	**8.2**	**3.9**

- En-De and En-ro are only trained using specified source and target languages
- En-Ne and En-Si, the pretraining is performed using mBART on 25 languages.
- mBART also generalizes well for the languages not seen in pretraining.

Results: mBART (only on source and target language) pretraining for unsupervised NMT

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv preprint arXiv:2001.08210, 2020.

When Unsupervised NMT does not work?

Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? arXiv preprint

Factors impacting the performance of Unsupervised NMT

- **Domain similarity**
  - Sensitive to domain mismatch
- **Dissimilar language pairs**
  - The similarity between language pairs helps the model in training good shared encoder
- **Initial model to start pretraining**
  - Good initializations leads to good performance in the finetuning phase
- **Unbalanced data size**
  - Not useful to use oversized data on one side
- **Quality of cross-lingual embeddings**
  - Initialization is done using cross-lingual embeddings

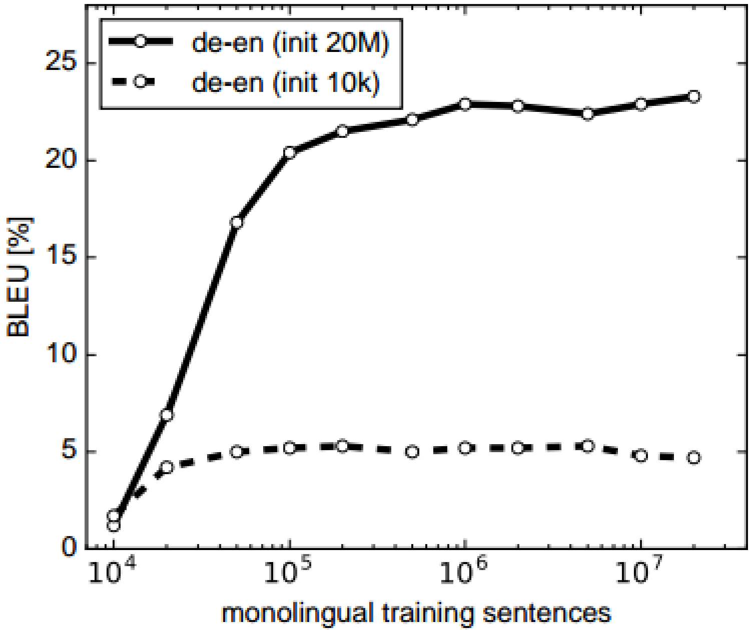
Domain similarity

Domain (en)	Domain (de/ru)	BLEU [%] de-en	en-de	ru-en	en-ru
Newswire	Newswire	23.3	19.9	11.9	9.3
	Politics	11.5	12.2	2.3	2.5
	Random	18.4	16.4	6.9	6.1

- Different distributions of the topics

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

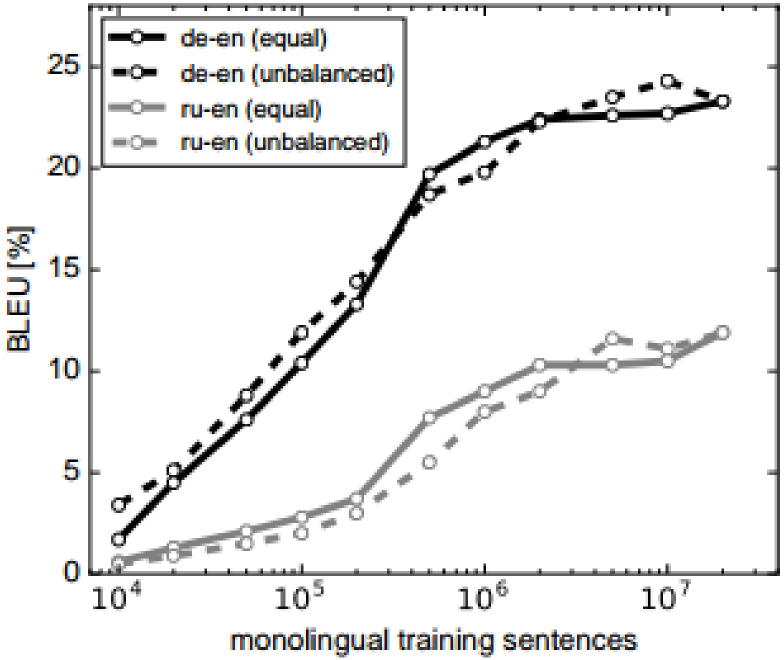
Initialization



- Good initializations leads to good performance in the fine-tuning phase
- Final model correlates well with the initialization quality

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

Unbalanced data size



Target side training data: 20M sentences

Solid line: target data has the same number of source and target sentences

- Not useful to use oversized data on one side

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

Quality of Cross-lingual Embeddings

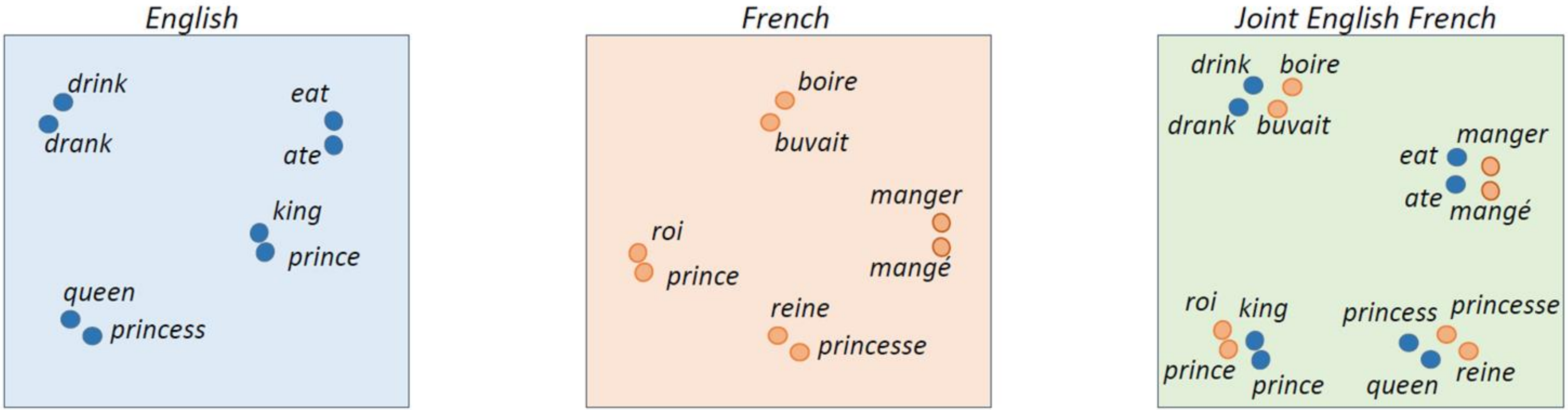
Cross-lingual Word Embeddings: Quality?

Unsupervised NMT [Lample et al 2018]

Pre-processing

1. Obtain cross-lingual embeddings either in an unsupervised manner or supervised manner
2. The pre-trained cross-lingual embeddings are not updated during training
3. Success of the approach relies on the quality of cross-lingual embeddings in addition to other factors like *language relatedness, etc*

Cross-lingual Representations

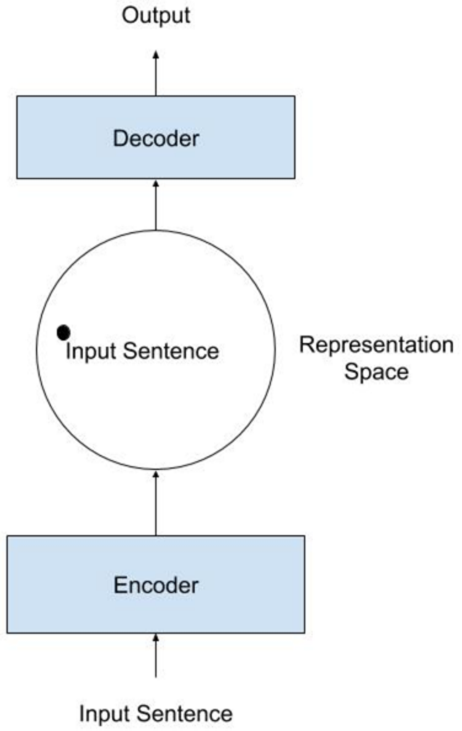


<u>Monolingual Word Representations</u>
(capture syntactic and semantic similarities between words)

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

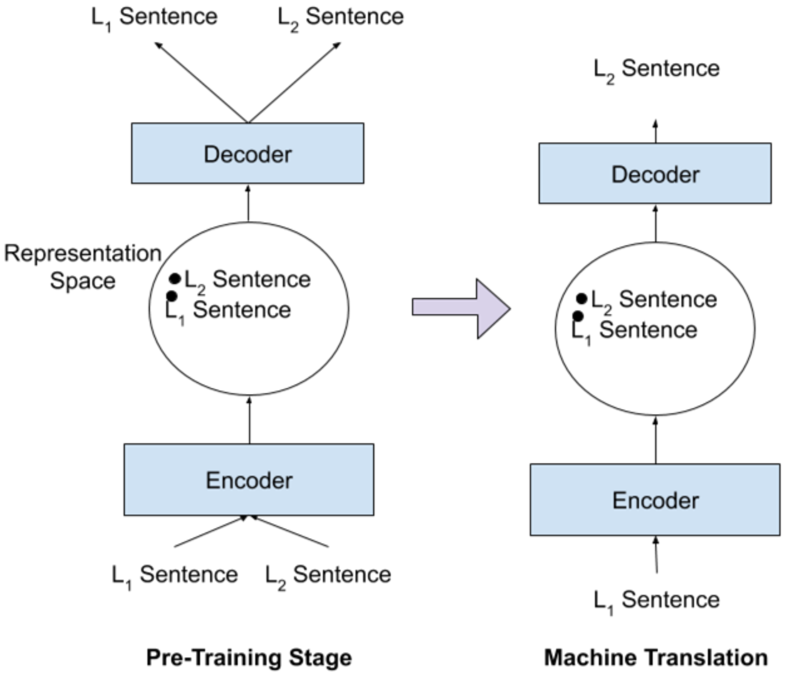
(Source: Khapra and Chandar, 2016)

Why is the Quality questioned?



Encode-Decode paradigm used for MT

Good Quality Cross-lingual Embeddings?



Encode-Decode paradigm used for MT

The ability of the encoder to learn better multilingual representations lies on the quality of cross-lingual embeddings

Quantitative Quality

Source - Target	GeoMM
En - Es	81.4
Es - En	85.5
En - Fr	82.1
Fr - En	84.1
En - De	74.7
De - En	76.7
En - Hi	41.5
Hi - En	54.8
En - Ta	31.9
Ta - En	38.7
En - Bn	36.7
Bn - En	42.7

Very low Precision@1 for Indic languages compared to the European language counterpart

Precision@1 for BLI task using GeoMM on MUSE dataset (*Jawapuria et.al 2019, Kakwani et.al 2020*)

Unsupervised NMT [Lample et al 2018]

Simple word-by-word translation using cross-lingual embeddings

	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70

Unsupervised NMT [Lample et al 2018]

Simple word-by-word translation using cross-lingual embeddings

Language Pair	BLEU Score
En → Fr	6.28
Fr → En	10.09
En → De	7.06
De → En	10.77
En → Hi	**1.2**
Hi → En	**2.1**

Credit: Tamali for the English-Hindi numbers

Cross-lingual Embedding Quality

1. Poor Cross-lingual Embeddings leads to diminished returns from U-NMT methods

Future Directions

1. Learn better cross-lingual embeddings between Indic languages and Indic to European languages
2. Majority of the NLP approaches operate at sub-word level
3. How to obtain cross-lingual embeddings at the sub-word level?

Unsupervised NMT for Indic languages

Initial Findings

Why Indic Languages?

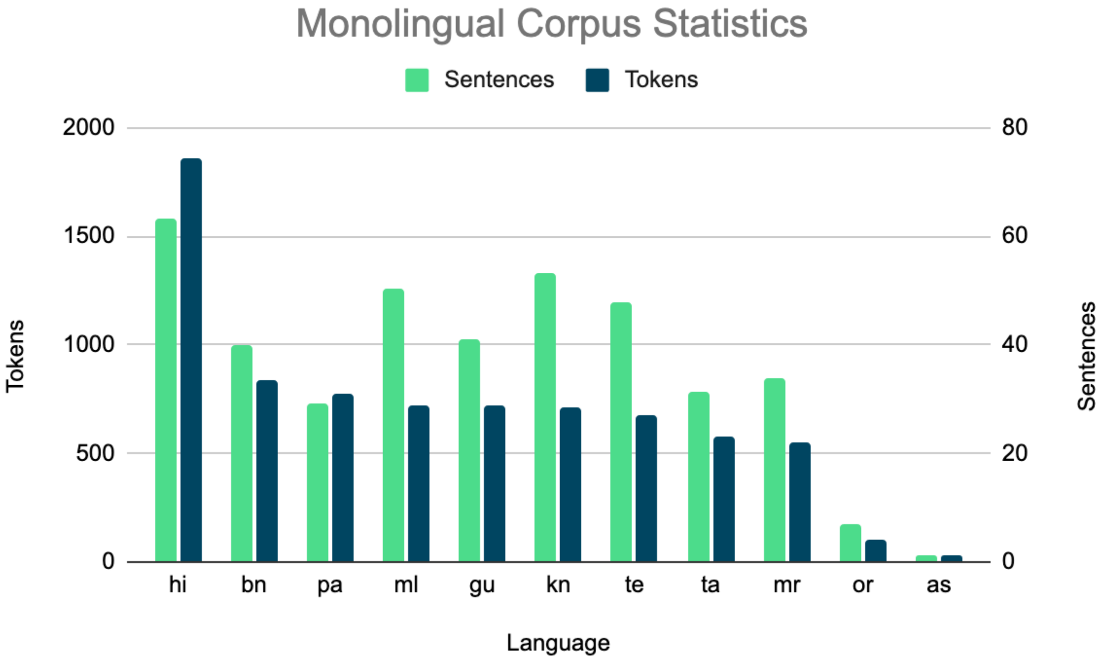
- A test-bed for research on multilinguality
- Spectrum of language similarity

	Bn	Gu	Hi	Mr	Pa	Ml	Ta	Te
Bn	-	19.51	29.45	11.39	2.45	1.05	0.34	0.78
Gu	13.9	-	51.75	20.14	4.46	1.06	0.3	1.22
Hi	12.76	31.47	-	15.22	4.43	0.78	0.21	0.95
Mr	11.81	29.31	36.42	-	3.4	0.62	0.27	0.92
Pa	4.26	10.88	17.79	5.71	-	0.22	0.16	0.4
Ml	1.19	1.7	2.04	0.67	0.14	-	0.72	2.48
Ta	0.43	0.54	0.62	0.33	0.11	0.8	-	0.25
Te	0.95	2.1	2.67	1.08	0.28	2.68	0.24	-

Percentage of words in the source language (row) which also appear in the target language (column) (transliterated to a common script) and having at least one common synset obtained from Indo-Wordnet (Bhattacharyya et.al 2010)

Why Indic Languages?

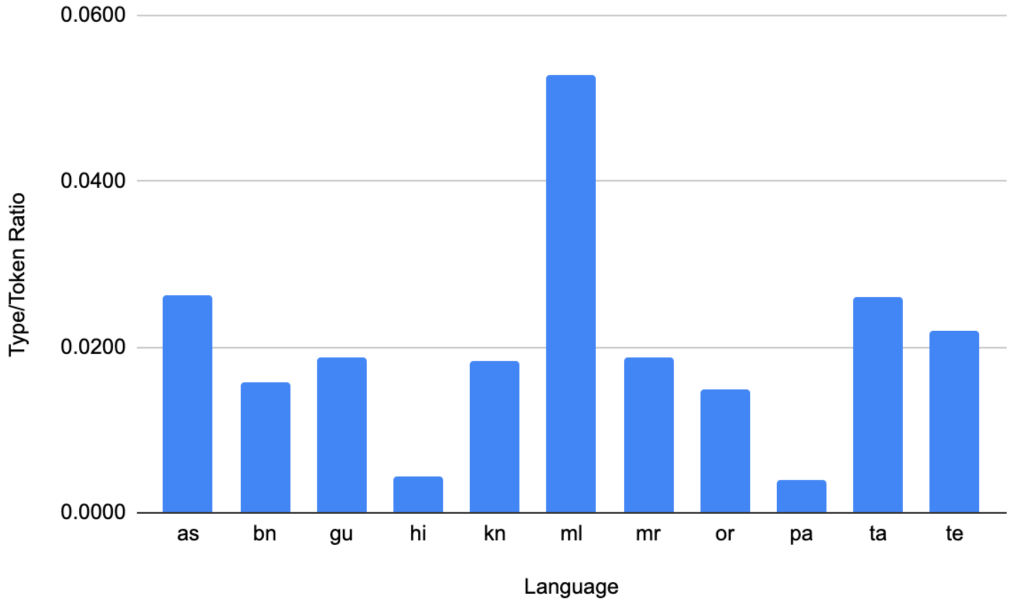
- Low-resourceness



Monolingual Corpus Statistics (in Millions) (Kunchukuttan et.al 2020)

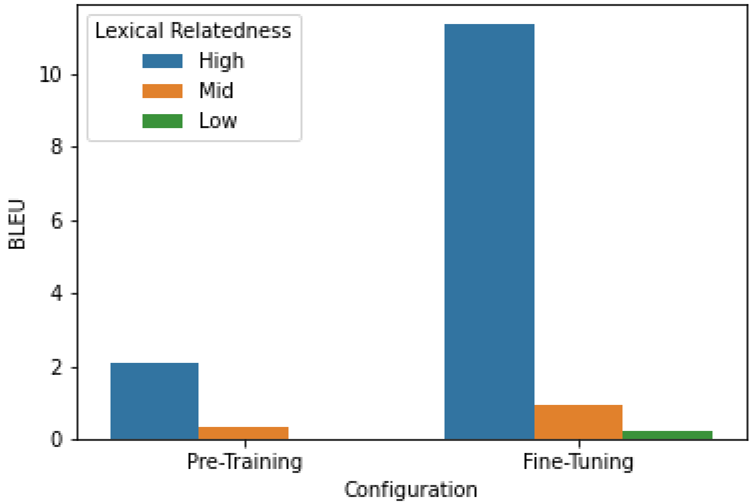
Why Indic Languages?

- Spectrum of morphological complexity

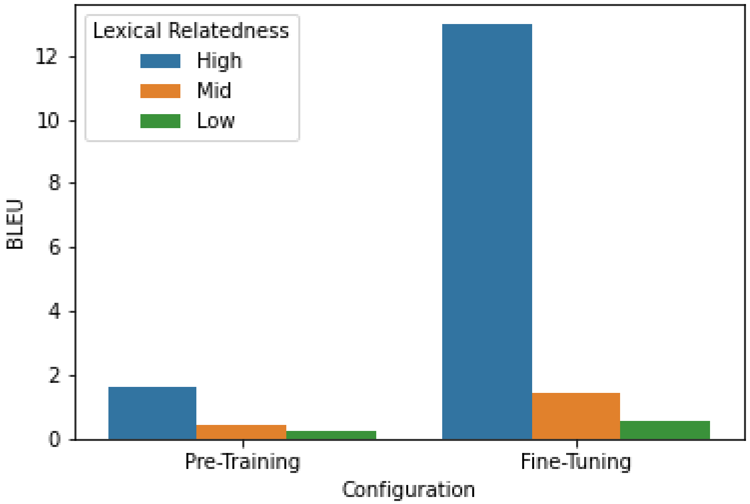


Type-Token Ratio calculated on AI4Bharat Corpus (Kunchukuttan et.al 2020)

U-NMT for Indic Languages: Results



Source → Target



Target → Source

Conclusions

1. Existing U-NMT models fail for Indic languages
2. For closely-related languages, we observe decent BLEU scores
3. Morphological richness adds more complexity to the model
4. Need **more research** focusing on **Indic languages**

Conclusions

Conclusion

- Paradigms of the MT task.
- Foundational concepts for the U-NMT paradigm.
- U-NMT approaches.
- Recent language modeling approaches.
- Results for Indian language pairs (related and unrelated languages).
- Need for further research in the area of U-NMT.

Future of U-NMT

1. U-NMT approaches have shown promising results for closely-related languages
2. U-NMT performs poor for distant languages
3. Better cross-lingual embeddings for distant languages.
4. Better cross-lingual language model pretraining for resource-scarce languages, disimilar languages, and dissimilar domans

Resources

- Resources can be found here

  www.cfilt.iitb.ac.in

- The tutorial slides will be uploaded here

  https://github.com/murthyrudra/unmt_tutorial_icon2020

References

- Sennrich R, Haddow B, Birch A. “**Improving Neural Machine Translation Models with Monolingual Data**.” In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2016 Aug (pp. 86-96).
- Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. "**Unsupervised Neural Machine Translation.**" In *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*.
- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "**Unsupervised Statistical Machine Translation.**" In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3632-3642. 2018.
- Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "**Iterative back-translation for neural machine translation**." In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.
- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "**An Effective Approach to Unsupervised Machine Translation.**" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 194-203. 2019.
- Lample, Guillaume, Alexis Conneau, and Ludovic Denoyer. "**Unsupervised Machine Translation Using Monolingual Corpora Only.**" In *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*.
- Lample, Guillaume, Myle Ott, Alexis Conneau, and Ludovic Denoyer. "**Phrase-Based & Neural Unsupervised Machine Translation.**" In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5039-5049. 2018.

References

- Yang, Zhen, Wei Chen, Feng Wang, and Bo Xu. "**Unsupervised Neural Machine Translation with Weight Sharing.**" In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 46-55. 2018.
- Wu, Jiawei, Xin Wang, and William Yang Wang. "**Extract and Edit: An Alternative to Back-Translation for Unsupervised Neural Machine Translation.**" In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1173-1183. 2019.
- Ren, Shuo, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. "**Unsupervised Neural Machine Translation with SMT as Posterior Regularization.**" (2019). In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 241-248. 2019.
- Conneau, Alexis, and Guillaume Lample. "**Cross-lingual language model pretraining.**" In *Advances in Neural Information Processing Systems*, pp. 7059-7069. 2019.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. "**MASS: Masked sequence to sequence pre-training for language generation**." In *International Conference on Machine Learning*, pp. 5926-5936. 2019.
- Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou and Shuai Ma. "**Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**".In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 770-779. 2019.

References

- Graça, Yunsu Kim Miguel, and Hermann Ney. "**When and Why is Unsupervised Neural Machine Translation Useless?.**" In *22nd Annual Conference of the European Association for Machine Translation*, p. 35.
- Marchisio, Kelly, Kevin Duh, and Philipp Koehn. **"When Does Unsupervised Machine Translation Work?.**" *arXiv preprint arXiv:2004.05516* (2020).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "**BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.**" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2020.
- Banerjee Tamali, Murthy V Rudra, Bhattacharyya Pushpak. **Ordering Matters: Word Ordering Aware Unsupervised NMT**. arXiv preprint arXiv:1911.01212. (2019).
- Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "**Word translation without parallel data.**" In *International Conference on Learning Representations*. 2018.
- Nima Pourdamghani and Kevin Knight. 2017. "**Deciphering Related Languages.**" In *Proceedings of EMNLP 2017*
- Iftekhar Naim, Parker Riley, and Daniel Gildea. 2018. "**Feature-Based Decipherment for Machine Translation. *Computational Linguistics*.**"
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). **Natural language processing (almost) from scratch**. Journal of Machine Learning Research
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). **A neural probabilistic language model**. Journal of Machine Learning Research

References

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). **Distributed representations of words and phrases and their compositionality**. Advances in Neural Information Processing Systems 26 (NIPS 2013)
- S Chandar AP, S Lauly, H Larochelle, M Khapra, B Ravindran, VC Raykar, **An autoencoder approach to learning bilingual word representations**. Advances in neural information processing systems 27, 1853-1861
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** In *NAACL-HLT (1)*. 2019.
- Kakwani, Divyanshu, Anoop Kunchukuttan, Satish Golla, N. C. Gokul, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. **iNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4948-4961. 2020.
- Jawanpuria, Pratik, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. **Learning multilingual word embeddings in latent metric space: a geometric approach.** *Transactions of the Association for Computational Linguistics* 7 (2019): 107-120.
- Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "**Iterative back-translation for neural machine translation.**" In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "**Improving Neural Machine Translation Models with Monolingual Data.**" In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

References

- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. **Multilingual denoising pre-training for neural machine translation**. arXiv preprint arXiv:2001.08210, 2020.

Backup Slides

TLM

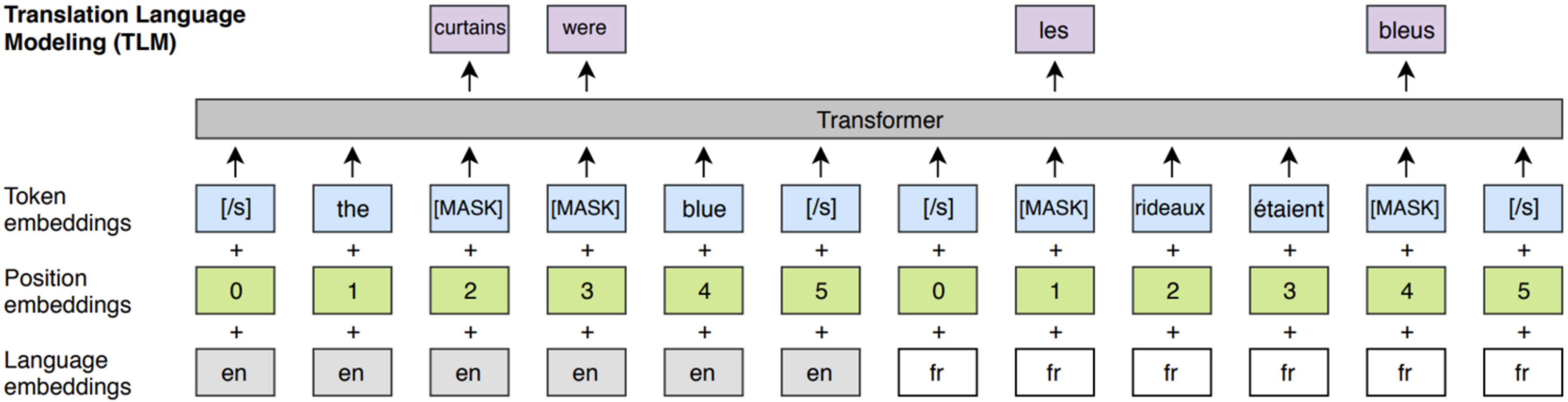
Translation Language Modelling

Cross-lingual Language Model Pretraining, ICLR, *Conneau et.al 2019*

TLM

- XLM objective uses monolingual corpora in all the languages considered
- Does XLM learn better multilingual representations?
  - XLM objective cannot take advantage of parallel corpora if available
  - XLM objective alone cannot guarantee that the model learns better multilingual representations

TLM (Conneau et.al 2019)



TLM (Conneau et.al 2019)

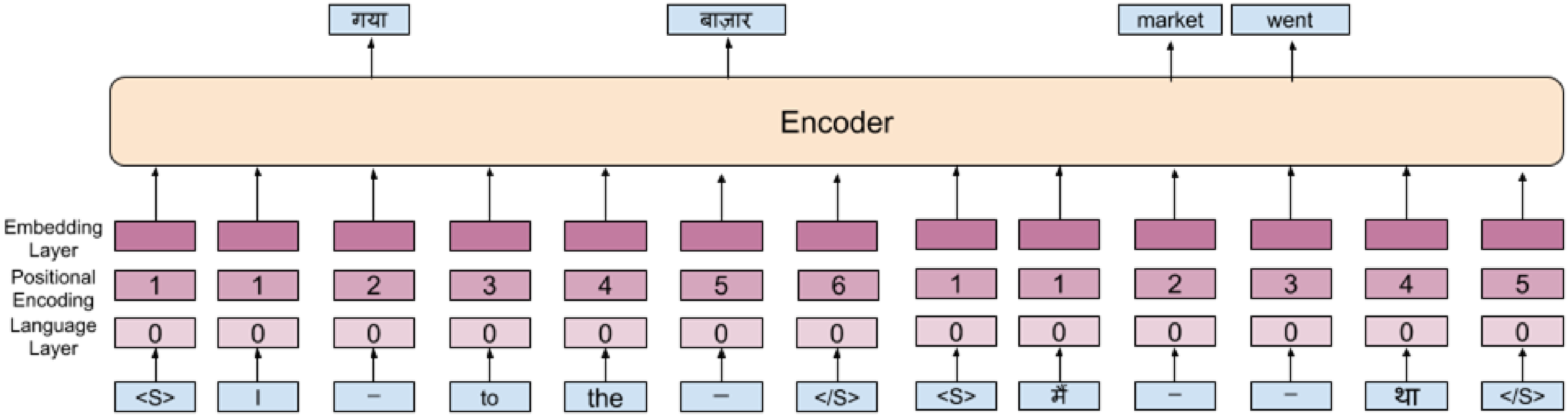
- In addition to access to monolingual corpus, we assume access to parallel corpus
- Given a parallel sentence,
    - The two sentences are concatenated and a special sentence delimiter is added to differentiate the two sentences
    - The positional information is reset to start from zero for the second language
    - The model can look at information from the context of either of the languages to predict the missing word

TLM (Conneau et.al 2019) : XNLI Results

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. [14]	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	<u>85.0</u>	<u>80.2</u>	<u>80.8</u>	<u>80.3</u>	<u>78.1</u>	<u>79.3</u>	<u>78.1</u>	<u>74.7</u>	<u>76.5</u>	<u>76.6</u>	<u>75.5</u>	<u>78.6</u>	<u>72.3</u>	<u>70.9</u>	63.2	<u>76.7</u>
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. [14]	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	<u>85.0</u>	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. [12]	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. [14]	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk [4]	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**<u>85.0</u>**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**<u>67.3</u>**	**75.1**

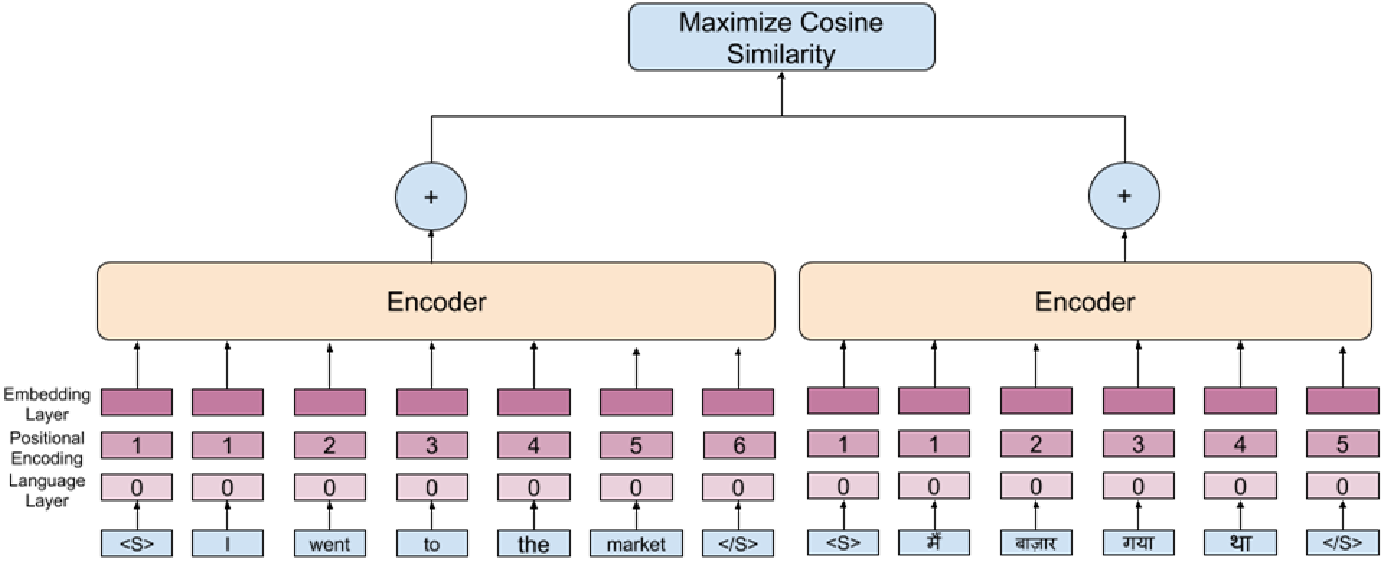
Extensions to TLM

- TLM model does not fully utilize the potential of parallel corpus
- Modify TLM objective to predict aligned words from the other language



Extensions to TLM

- Maximize the cosine similarity between the encoder representation of the two sentences



Challenges in Indic Languages?

Original Sentence	Comments	Google Translate[30 Nov,2020]
ನಾನು ಹೇಳುವುದನ್ನು ಸರಿಯಾಗಿ ಕೇಳಿಸಿಕೋ nAnu heLuvudannu sariyAgi keLisiko Me telling correctly listen	Literary Language	Listen to me correctly
ನಾನು ಹೇಳೋದನ್ನ ಸರ್ಯಾಗಿ ಕೇಳ್ಸ್ಕೊ nAnu heLodanna saryAgi keLsko	Spoken Language	I am Sergio Katsko of Noodon
ಊಟ ಮಾಡಿಕೊಂಡು ಹೋಗು UTa mADikoMDu hogu Lunch have go	Literary Language	Go Have lunch (Go after having lunch)
ಊಟ ಮಾಡ್ಕೊಂಡು ಹೋಗು UTa mADkoMDu hogu	Spoken Language	Modify the meal

Phenomenon similar to Schwa Deletion in Literary language and Spoken language

Why Indic Languages?

Original Sentence	Comments	Google Translate
ವಂಚಕಾಸುರರನ್ನೊದ್ದೋಡಿಸಿರುವವನಾರೆಂದೇನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ?	Maximum Sandhi transformation	Do you know anyone who has cheated?
ವಂಚಕ ಅಸುರರನ್ನು ಒದ್ದು ಓಡಿಸಿ ಇರುವ ಅವನು ಯಾರು ಎಂದು ಏನು ಆದರು ನಿಮಗೆ ತಿಳಿದು ಇದೆಯೇ ?	No Sandhi transformation	Do you know who became the one who drove out the crafty demons?
ವಂಚಕಾಸುರರನ್ನು ಒದ್ದೋಡಿಸಿರುವವನು ಯಾರೆಂದು ಏನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ ? Crooked demons one who kicked them away who is you know	Normal Usage	Do you know who is the one who kicked the crooks?

Components of U-MT

- Suitable initialization of the translation models: This helps the model to jump-start the process.
- Language modeling: This helps the model to encode and generate sentences.
- Iterative back-translation: It bridges the gap between encoder representation of a word in source and target languages.

Adding subword information

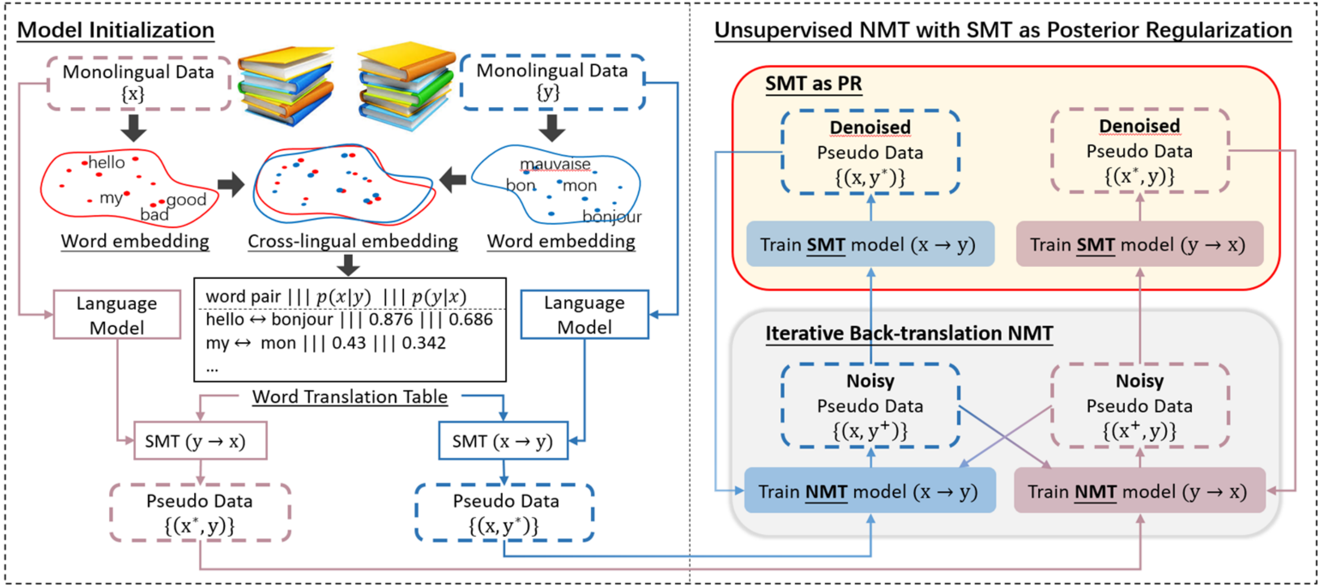
- We want to favor translation candidates that are similar at the character level.
- Additional weights are added to initial phrase-table like lexical weightings.
  - Unlike lexical weightings it use **a character-level similarity function** instead of word translation probabilities.

$$\mathrm{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \mathrm{sim}(\bar{f}_i, \bar{e}_j)\right)$$

$$\mathrm{sim}(f,e) = 1 - \frac{\mathrm{lev}(f,e)}{\max(\mathrm{len}(f), \mathrm{len}(e))}$$

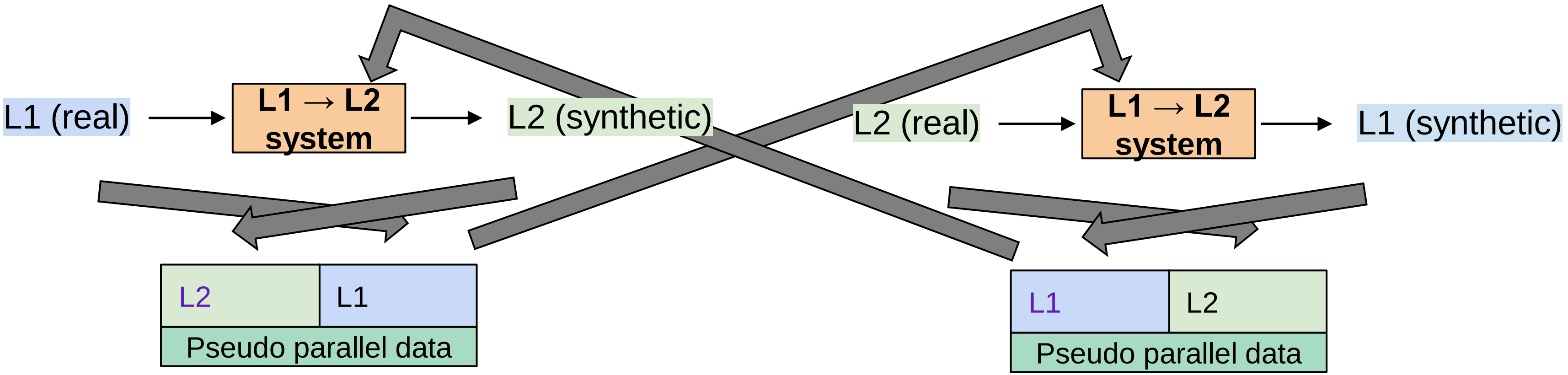
USMT as Posterior Regularization

- USMT initialisation.
- UNMT backtranslation training with SMT as Posterior Regularization.
  - Posterior Regularization: An SMT system to filter out noises using phrase table. It eliminates the infrequent and bad patterns generated in the back-translation iterations of NMT



Iterative refinement

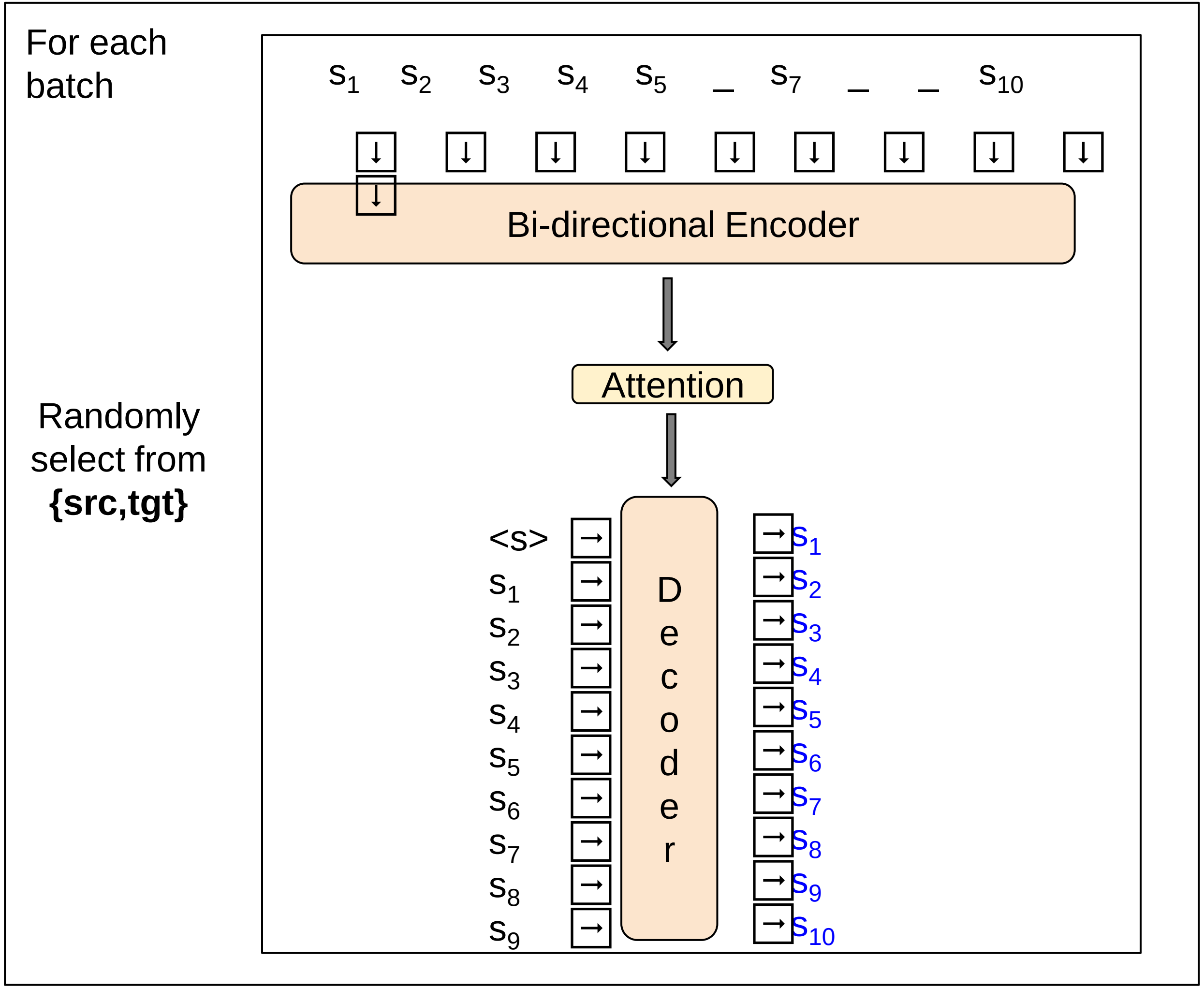
- Generate a synthetic parallel corpus by translating the monolingual corpus with the initial system L1→L2, and train and tune SMT system L2→L1.
  - To accelerate our experiments, use a random subset of 2 million sentences from each monolingual corpus for training.
  - Reuse the original language model, which is trained in the full corpus.
- The process is repeated iteratively until some convergence criterion is met.



BART Pretraining

- Trained by
  - Corrupting text with an arbitrary noising function
  - Learning a model to reconstruct the original text.
- Denoising full text

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL 2020)*



BART pretraining (noising steps) (Lewis et al. 2020)

A B C . D E .

Original document

Token Masking	A _ C . _ E .
Token deletion	A . C . E .
Text infilling	A _ . D _ E .
Sentence permutation	D E . A B C.
Document rotation	C . D E . A B .