Unsupervised Neural Machine Translation

Tutorial @ ICON-2020
IIT Patna

Prof. Pushpak Bhattacharyya
Diptesh Kanojia
Jyotsana Khatri
Tamali Banerjee
Rudra Murthy
Paradigms of Machine Translation

Acknowledgement: Numerous PhD, masters and UG students and research staff working on MT with me since 2000
Perspective
NLP: a useful view

- Increased Complexity Of Processing
- Morphology
- POS tagging
- Chunking
- Parsing
- Semantics
- Discourse and Coreference

Problem
- Semantics
- Parsing
- Part of Speech Tagging
- Morph Analysis

Language
- Hindi
- Marathi
- French
- English

Algorithm
- CRF
- HMM
- MEMM

NLP Trinity
Machine Translation: Translating from one language to another by computer

Knowledge Based: Essence - Analysis

Data Driven

Similarity based: Essence - Analogy

Probability based: Essence - Alignment

Distributed Representation based: Essence - Attention?

RBMT - rule based MT
EBMT - example based MT
SMT - Statistical MT
NMT - Neural MT
Today’s Ruling Paradigm: NMT which is data intensive

Essential Elements of MT Paradigms

- **Analysis** in RBMT
- **Alignment** in SMT
- **Analogy** in EBMT
- **Attention** 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

<table>
<thead>
<tr>
<th>Hindi; Indo Aryan</th>
<th>Tamil; Dravidian</th>
</tr>
</thead>
<tbody>
<tr>
<td>राजा को नमन करो</td>
<td>अरங்கர வணங்கு</td>
</tr>
<tr>
<td>raajaa ko naman karo</td>
<td>aracarai vanaNku</td>
</tr>
<tr>
<td>HG: king to obeisance do</td>
<td>king_to_obeisance_do</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English; Indo-Aryan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give obeisance to the king</td>
</tr>
<tr>
<td>king_to_obeisance_do</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marathi; Indo Aryan</th>
</tr>
</thead>
<tbody>
<tr>
<td>राजाला नमन करा</td>
</tr>
<tr>
<td>raajaalaa naman karaa</td>
</tr>
<tr>
<td>king_to_obeisance_do</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Manipuri; Tibeto Burman</th>
</tr>
</thead>
<tbody>
<tr>
<td>विंशोबू खौरम्मू</td>
</tr>
<tr>
<td>niNgthoubu khoirammu</td>
</tr>
<tr>
<td>king_to_obeisance_do</td>
</tr>
</tbody>
</table>
transfer amongst different language families

<table>
<thead>
<tr>
<th>Language</th>
<th>Inflected Verb/Inflected verb complex</th>
<th>Inflected Noun/Inflected Noun chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td><em>give obeisance</em></td>
<td><em>To the king</em></td>
</tr>
<tr>
<td>Hindi</td>
<td><em>naman karo</em></td>
<td><em>raajaa ko</em></td>
</tr>
<tr>
<td>Marathi</td>
<td><em>naman karaa</em></td>
<td><em>raajaalaa</em></td>
</tr>
<tr>
<td>Tamil</td>
<td><em>vanaNku</em></td>
<td><em>aracarai</em></td>
</tr>
<tr>
<td>Manipuri</td>
<td><em>Khoirammu</em></td>
<td><em>niNgthoubu</em></td>
</tr>
</tbody>
</table>
English parse tree

Transfer rules:

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

Source Text

Morphological Analyzer

POS Tagger

Chunker

Vibhakti Computation

Name Entity Recognizer

Word Sense Disambiguation

Transfer

Target Text

Word Generator

Interchunk

Intrachunk

Agreement Feature

Lexical Transfer
Foundation

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

$$
\hat{e} = \arg\max_{e \in e^*} p(e|f) = \arg\max_{e \in e^*} p(f|e)p(e)
$$

- 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 $w_1w_2......w_n$ can be written as
  $$Pr(w_1w_2......w_n) = Pr(w_1) \times Pr(w_2|w_1) \times ... \times Pr(w_n|w_1w_2...w_{n-1})$$
- Here $Pr(w_n|w_1w_2...w_{n-1})$ is the probability that word $w_n$ follows word string $w_1w_2...w_{n-1}$.
  - N-gram model probability
- Trigram model probability calculation
  $$p(w_3|w_1w_2) = \frac{count(w_1w_2w_3)}{count(w_1w_2)}$$
SMT: Translation Model

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

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

- 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

$$\Pr(f|e) = \sum_a \Pr(f, a|e)$$

- Sentence level
- Word level
Alignment

- If the string, $e = e_1 e_2 \ldots e_l$, has $l$ words, and the string, $f = f_1 f_2 \ldots f_m$, has $m$ words,
- then the alignment, $a$, can be represented by a series, $a_1^m = a_1 a_2 \ldots a_m$, 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
  - $a_j = i$, and
  - if it is not connected to any English word, then $a_j = 0$
Example of alignment

English: *Ram went to school*

Hindi: *raam paathashaalaa gayaa*
Translation Model: Exact expression

\[
Pr(f, a | e) = Pr(m | e) \prod_{j=1}^{m} Pr(a_j | a_{j-1}^{j-1}, f_{j-1}^{j-1}, m, e) Pr(f_j | a_1^{j}, f_1^{j-1}, m, e)
\]

- 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

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

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

\[ = \sum_m \Pr(m \mid e) \prod_{j=1}^m \Pr(f_j, a_j \mid a_{j-1}^{j-1}, f_{j-1}^{j-1}, m, e) \]

\[ = \sum_m \Pr(m \mid e) \prod_{j=1}^m \Pr(a_j \mid a_{j-1}^{j-1}, f_{j-1}^{j-1}, m, e) \Pr(f_j \mid a_j^i, f_{j-1}^{j-1}, m, e) \]

\( m \) is fixed for a particular \( f \), hence

\[ \Pr(f, a, m \mid e) = \Pr(m \mid e) \prod_{j=1}^m \Pr(a_j \mid a_{j-1}^{j-1}, f_{j-1}^{j-1}, m, e) \Pr(f_j \mid a_j^i, f_{j-1}^{j-1}, m, e) \]
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 $\rightarrow$ Phrase alignment (Koehn et al, 2003)
- Word alignment $\rightarrow$ 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**
(1) three rabbits
   a     b
(2) rabbits of Grenoble
   b     c     d

**French**
(1) trois lapins
   w     x
(2) lapins de Grenoble
   x     y     z
Initial Probabilities:
each cell denotes $t(a \leftrightarrow w)$, $t(a \leftrightarrow x)$ etc.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>x</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>y</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>z</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
</tbody>
</table>
Example of expected count

\[ C[w \leftarrow \rightarrow a; (a \ b) \leftarrow \rightarrow (w \ x)] \]

\[
\begin{align*}
t(w \leftarrow \rightarrow a) \\
= & \quad \frac{1}{4} + \frac{1}{4} \\
= & \quad \frac{1}{2}
\end{align*}
\]

\[
\begin{align*}
&= \quad \frac{1}{4} + \frac{1}{4} \\
&= \quad \frac{1}{2}
\end{align*}
\]
"counts"

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>w</strong></td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>y</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>z</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>w</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td><strong>y</strong></td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td><strong>z</strong></td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
Revised probability: example

\[ t_{\text{revised}}(a \leftrightarrow w) \]

\[ \frac{1}{2} \]

\[ = \frac{1}{2} + \frac{1}{2} + 0 + 0 + 0 + 0 \]

\[ \left( \frac{1}{2} + 1 \right)_{(a \ b)} \leftrightarrow_{(w \ x)} + (0 + 0 + 0 + 0)_{(b \ c \ d)} \leftrightarrow_{(x \ y \ z)} \]
## Revised probabilities table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1/4</td>
<td>5/12</td>
<td>1/6</td>
<td>1/6</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
"revised counts"

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>1/2</td>
<td>3/8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1/2</td>
<td>5/8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>5/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
Re-Revised probabilities table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>3/16</td>
<td>85/144</td>
<td>1/9</td>
<td>1/9</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Continue until convergence; notice that (b,x) binding gets progressively stronger; b=rabbits, x=lapins
Derivation of EM based Alignment Expressions

\[ V_E = \text{vocabulary of language } L_1 \text{ (Say English)} \]
\[ V_F = \text{vocabulary of language } L_2 \text{ (Say Hindi)} \]

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.

jise hum gulab kahate hai, aur bhi kisi naam se uski kushbu saman mitha hogii

F2 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

<table>
<thead>
<tr>
<th>$V_E$</th>
<th>$V_F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>what, is, in, a, name, that, which, we, call, rose, by, any, other, will, smell, as, sweet</td>
<td>naam, meM, kya, hai, jise, ham, gulab, kahte, aur, bhi, kisi, bhi, uski, khushbu, saman, mitha, hogii</td>
</tr>
</tbody>
</table>
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,

$$e^1_1, e^1_2, ..., e^1_{l^1} \Leftrightarrow f^1_1, f^1_2, ..., f^1_{m^1}$$
$$e^2_1, e^2_2, ..., e^2_{l^2} \Leftrightarrow f^2_1, f^2_2, ..., f^2_{m^2}$$

......

$$e^s_1, e^s_2, ..., e^s_{l^s} \Leftrightarrow f^s_1, f^s_2, ..., f^s_{m^s}$$

......

$$e^s_1, e^s_2, ..., e^s_{l^s} \Leftrightarrow f^s_1, f^s_2, ..., f^s_{m^s}$$

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

Data Likelihood $L(D; \Theta)$:

$$L(D; \Theta) = \prod_{s=1}^{S} \prod_{p=1}^{l^s} \prod_{q=1}^{m^s} \left( P_{index_E(e^s_p),index_F(f^s_q)} \right)^{z^s_{pq}}$$

Data Log-Likelihood $LL(D; \Theta)$:

$$LL(D; \Theta) = \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} z^s_{pq} \log \left( P_{index_E(e^s_p),index_F(f^s_q)} \right)$$

Expected value of Data Log-Likelihood $E(LL(D; \Theta))$:

$$E(LL(D; \Theta)) = \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} E(z^s_{pq}) \log \left( P_{index_E(e^s_p),index_F(f^s_q)} \right)$$
Constraint and Lagrangian

\[ \sum_{j=1}^{\left|V_F\right|} P_{i,j} = 1, \forall i \]

\[
\sum_{s=1}^{S} \sum_{p=1}^{l_s^s} \sum_{q=1}^{m_s^s} E(z_{pq}^s) \log\left(\frac{P_{\text{index}_B(f_q^s),\text{index}_F(f_q^s)}}{P_{\text{index}_B(f_q^s),\text{index}_F(f_q^s)}}\right) - \sum_{i=1}^{\left|V_E\right|} \lambda_i \left(\sum_{j=1}^{\left|V_F\right|} P_{i,j} - 1\right)
\]
Differentiating wrt $P_{ij}$

$$\sum_{s=1}^{S} \sum_{l^s} \sum_{m^s} \delta_{\text{index}_E(e^s_p),i} \delta_{\text{index}_F(f^s_q),j} \left( \frac{E(z^s_{pq})}{P_{i,j}} \right) - \lambda_i = 0$$

$$P_{i,j} = \frac{1}{\lambda_i} \sum_{s=1}^{S} \sum_{l^s} \sum_{m^s} \delta_{\text{index}_E(e^s_p),i} \delta_{\text{index}_F(f^s_q),j} E(z^s_{pq})$$

$$\sum_{j=1}^{|V_F|} P_{i,j} = 1 = \sum_{j=1}^{|V_F|} \frac{1}{\lambda_i} \sum_{s=1}^{S} \sum_{l^s} \sum_{m^s} \delta_{\text{index}_E(e^s_p),i} \delta_{\text{index}_F(f^s_q),j} E(z^s_{pq})$$
Final E and M steps

M-step

\[
  P_{i,j} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{l_s} \sum_{q=1}^{m_s} \delta_{\text{index}_E(e_p^s),i} \delta_{\text{index}_F(f_q^s),j} E(z_{pq}^s)}{\sum_{j=1}^{|V_F|} \sum_{s=1}^{S} \sum_{p=1}^{l_s} \sum_{q=1}^{m_s} \delta_{\text{index}_E(e_p^s),i} \delta_{\text{index}_F(f_q^s),j} E(z_{pq}^s)}, \forall i, j
\]

E-step

\[
  E(z_{pq}^s) = \frac{P_{\text{index}_E(e_p^s),\text{index}_F(f_q^s)}}{\sum_{q_1=1}^{m_s} P_{\text{index}_E(e_p^s),\text{index}_F(f_{q_1}^s)}}, \forall s, p, q
\]
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, Rekjavik, Iceland, 26-31 May, 2014

Kunchukuttan & Bhattacharyya (EMNLP 2016)
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
“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)
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

\[
BP = \begin{cases} 
\frac{1}{e^{(1-r/c)}} & \text{if } c > r \\
1 & \text{if } c \leq r 
\end{cases}
\]

Precision -> Modified n-gram precision

\[
p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Count}_{\text{clip}}(n-\text{gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n-\text{gram}' \in C'} \text{Count}(n-\text{gram}')} 
\]

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

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)

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

<table>
<thead>
<tr>
<th>SL No.</th>
<th>Feature</th>
<th>Value</th>
<th>Similarity function s(·)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Length</td>
<td>Integer</td>
<td>equality</td>
</tr>
<tr>
<td>2</td>
<td>Active/Passive</td>
<td>1 (active) / 0 (passive)</td>
<td>equality</td>
</tr>
<tr>
<td>3</td>
<td>Parse tree</td>
<td>–</td>
<td>Tree similarity between</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>two parse trees</td>
</tr>
<tr>
<td>4</td>
<td>Concatenation of vectors of words forming the sentence</td>
<td>Vector of Boolean/real values</td>
<td>Cosine similarity</td>
</tr>
<tr>
<td>5</td>
<td>Bag of words forming the sentence</td>
<td>Set</td>
<td>Dice/Jaccard and such other similarity measures</td>
</tr>
<tr>
<td>6</td>
<td>Position of nouns of the sentence in the wordnet hypernym hierarchy</td>
<td>A function combining the information content of the individual nouns</td>
<td>equality</td>
</tr>
<tr>
<td>7</td>
<td>Position of the two main verbs of the sentence in VerbOcean²</td>
<td>“distance”^1</td>
<td>A rule that says similar or dissimilar, depending on the distance being within a threshold or not</td>
</tr>
<tr>
<td>8</td>
<td>main verb, its type and argument frame</td>
<td>A slot-filler structure for each sentence</td>
<td>Equality or subset-check on the slots and their fillers</td>
</tr>
<tr>
<td>9</td>
<td>Frame semantic representation of the sentence as per FrameNet⁴</td>
<td>Slot-filler structure</td>
<td>Equality or subset-check on the slots and their fillers</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Language pair</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi - Mr</td>
<td>31.25</td>
</tr>
<tr>
<td>Hi - Pa</td>
<td>63.38</td>
</tr>
<tr>
<td>Pa - Hi</td>
<td>68.31</td>
</tr>
<tr>
<td>Hi - Gu</td>
<td>49.98</td>
</tr>
<tr>
<td>Gu - Hi</td>
<td>53.22 (↑ from 53.09 from SMT)</td>
</tr>
</tbody>
</table>
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 होटेलमध्ये एक संग्रहालय आहे/ hyaa hotel madhye ek saMgraahaalay aahe
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 “lal … aahe”

If syntactic subject is inanimate
Then
“have” should translate to “madhye … aahe”
$X$ have $Y$ →

$X_{\text{kade}}$ $Y$ aahe /

$X_{\text{laa}}$ $Y$ aahe /

$X_{\text{madhye}}$ $Y$ aahe
• 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 ←→ hospital madhye aahe
  hospital <cm> has
“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 chalaaing”)

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

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

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 $\rightarrow$ will+eat” AND “jaa+rahaa_hE $\rightarrow$ is+going”
    – Produce “khaa+rahaa_hE $\rightarrow$ 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**

<table>
<thead>
<tr>
<th>Source</th>
<th>pan</th>
<th>hin</th>
<th>guj</th>
<th>ben</th>
<th>mar</th>
<th>kok</th>
<th>tel</th>
<th>tam</th>
<th>mar</th>
</tr>
</thead>
<tbody>
<tr>
<td>pan</td>
<td>71.40</td>
<td>64.56</td>
<td>44.52</td>
<td>27.97</td>
<td>36.50</td>
<td>32.44</td>
<td>26.89</td>
<td>19.53</td>
<td>16.61</td>
</tr>
<tr>
<td>hin</td>
<td></td>
<td>49.14</td>
<td>29.70</td>
<td>18.37</td>
<td>24.88</td>
<td>21.33</td>
<td>18.20</td>
<td>14.85</td>
<td>13.73</td>
</tr>
<tr>
<td>guj</td>
<td></td>
<td></td>
<td>23.88</td>
<td>14.35</td>
<td>21.33</td>
<td>18.20</td>
<td>16.15</td>
<td>12.44</td>
<td>11.18</td>
</tr>
<tr>
<td>ben</td>
<td></td>
<td></td>
<td></td>
<td>13.84</td>
<td>16.15</td>
<td>15.99</td>
<td>10.93</td>
<td>10.67</td>
<td>11.20</td>
</tr>
<tr>
<td>mar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.64</td>
<td>10.92</td>
<td>7.93</td>
<td>7.74</td>
<td>7.99</td>
</tr>
<tr>
<td>kok</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.93</td>
<td>7.93</td>
<td>7.74</td>
<td>7.99</td>
</tr>
<tr>
<td>tel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.74</td>
<td>7.74</td>
<td>7.74</td>
</tr>
<tr>
<td>tam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.74</td>
<td>7.74</td>
</tr>
<tr>
<td>mar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.74</td>
</tr>
</tbody>
</table>

**% improvement over word level scores**

<table>
<thead>
<tr>
<th>Source</th>
<th>pan</th>
<th>hin</th>
<th>guj</th>
<th>ben</th>
<th>mar</th>
<th>kok</th>
<th>tel</th>
<th>tam</th>
<th>mar</th>
</tr>
</thead>
<tbody>
<tr>
<td>pan</td>
<td>2.04</td>
<td>11.61</td>
<td>10.35</td>
<td>12.77</td>
<td>10.75</td>
<td>8.61</td>
<td>15.72</td>
<td>21.74</td>
<td></td>
</tr>
<tr>
<td>hin</td>
<td>1.94</td>
<td>10.15</td>
<td>7.45</td>
<td>6.31</td>
<td>2.75</td>
<td>8.77</td>
<td>12.72</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>guj</td>
<td>7.95</td>
<td>7.32</td>
<td>8.86</td>
<td>8.71</td>
<td>7.93</td>
<td>6.38</td>
<td>11.28</td>
<td>22.89</td>
<td></td>
</tr>
<tr>
<td>ben</td>
<td>1.19</td>
<td>7.75</td>
<td>7.43</td>
<td>5.00</td>
<td>8.53</td>
<td>9.18</td>
<td>11.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mar</td>
<td>13.85</td>
<td>9.99</td>
<td>14.79</td>
<td>12.11</td>
<td>5.60</td>
<td>12.64</td>
<td>13.16</td>
<td>18.39</td>
<td></td>
</tr>
<tr>
<td>kok</td>
<td>10.45</td>
<td>3.83</td>
<td>10.50</td>
<td>9.50</td>
<td>6.84</td>
<td>8.77</td>
<td>18.22</td>
<td>21.61</td>
<td></td>
</tr>
<tr>
<td>tel</td>
<td>10.48</td>
<td>10.21</td>
<td>18.91</td>
<td>17.04</td>
<td>20.88</td>
<td>15.17</td>
<td>14.57</td>
<td>14.76</td>
<td></td>
</tr>
<tr>
<td>tam</td>
<td>7.25</td>
<td>9.03</td>
<td>14.11</td>
<td>13.71</td>
<td>15.83</td>
<td>14.56</td>
<td>7.56</td>
<td>23.64</td>
<td></td>
</tr>
<tr>
<td>mar</td>
<td>27.38</td>
<td>27.64</td>
<td>29.74</td>
<td>32.53</td>
<td>27.05</td>
<td>30.54</td>
<td>21.08</td>
<td>22.92</td>
<td></td>
</tr>
</tbody>
</table>
Factor based SMT

 Semantic relations + Suffixes → Case Markers + inflections

\[ I \text{ ate mangoes} \]

\[ I \{<\text{agt}\} \text{ ate } \{\text{eat@past}\} \text{ mangoes } \{<\text{obj}\} \]

\[ I \{<\text{agt}\} \text{ mangoes } \{<\text{obj.@pl}\} \{\text{eat@past}\} \]

\[ mei_{-}ne \text{ aam } khaa_{-}yaa \]
Our Factorization based on Koehn and Hoang (2007)

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

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

3. a lemma + suffix to surface form generation factor (लडक + ᴗ (ladak + e) → लडके (ladake))
Experiment: Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>#sentences</th>
<th>#words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>12868</td>
<td>316508</td>
</tr>
<tr>
<td>Tuning</td>
<td>600</td>
<td>15279</td>
</tr>
<tr>
<td>Test</td>
<td>400</td>
<td>8557</td>
</tr>
</tbody>
</table>
Results: The impact of suffix and semantic factors

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (surface)</td>
<td>24.32</td>
<td>5.85</td>
</tr>
<tr>
<td>lemma + suffix</td>
<td>25.16</td>
<td>5.87</td>
</tr>
<tr>
<td>lemma + suffix + unl</td>
<td>27.79</td>
<td>6.05</td>
</tr>
<tr>
<td>lemma + suffix + stanford</td>
<td>28.21</td>
<td>5.99</td>
</tr>
</tbody>
</table>
**Results: The impact of reordering and semantic relations**

<table>
<thead>
<tr>
<th>Model</th>
<th>Reordering</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>surface</td>
<td>distortion</td>
<td>24.42</td>
<td>5.85</td>
</tr>
<tr>
<td>surface</td>
<td>lexicalized</td>
<td>28.75</td>
<td>6.19</td>
</tr>
<tr>
<td>surface</td>
<td>syntactic</td>
<td>31.57</td>
<td>6.40</td>
</tr>
<tr>
<td>lemma + suffix + stanford</td>
<td>syntactic</td>
<td>31.49</td>
<td>6.34</td>
</tr>
</tbody>
</table>

Table 5: Results: The impact of reordering and semantic relations
Subjective Evaluation: The impact of reordering and semantic relations

<table>
<thead>
<tr>
<th>Model</th>
<th>Reordering</th>
<th>Fluency</th>
<th>Adequacy</th>
<th>#errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>surface</td>
<td>lexicalized</td>
<td>2.14</td>
<td>2.26</td>
<td>2.16</td>
</tr>
<tr>
<td>surface</td>
<td>syntactic</td>
<td>2.6</td>
<td>2.71</td>
<td>1.79</td>
</tr>
<tr>
<td>lemma + suffix + stanford</td>
<td>syntactic</td>
<td>2.88</td>
<td>2.82</td>
<td>1.44</td>
</tr>
</tbody>
</table>
Cooperative NLP: Pivot Based MT

Triangulation

Water (English)

paanii (Hindi)

Jal (Bengali)
$L_1 \rightarrow \text{bridge} \rightarrow L_2$ (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

\[ e_{\text{best}} = \arg \max_e p(e|f) \]
\[ = \arg \max_e p(f|e)p_{LM}(e) \]

Where \( p(f|e) \) is given by:

\[ p(f|e) = p(I|\bar{e}) = \prod_{i=1}^{I} \phi(f_i|\bar{e}_i) d(a_i - b_{i-1}) p_w(f_i|\bar{e}_i, a) \gamma \]

\[ \phi(f_i|\bar{e}_i) = \sum_{\bar{p}_i} \phi(f_i|\bar{p}_i) \phi(\bar{p}_i|\bar{e}_i) \]

\[ p_w(f_i|\bar{e}_i, a) = \prod_{l=1}^{n} \frac{1}{|m|(l,m) \in a} \sum_{v(l,m) \in a} w(f_i|e_v) \]
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

<table>
<thead>
<tr>
<th>French</th>
<th>Creole</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>avion</td>
<td>Avion</td>
<td>aeroplane</td>
</tr>
<tr>
<td>bon</td>
<td>Bon</td>
<td>good</td>
</tr>
<tr>
<td>gaz</td>
<td>Gaz</td>
<td>gas</td>
</tr>
<tr>
<td>bref</td>
<td>bref</td>
<td>brief</td>
</tr>
<tr>
<td>pion</td>
<td>pion</td>
<td>pawn</td>
</tr>
</tbody>
</table>
## Experiment on MCR→FR→E

<table>
<thead>
<tr>
<th>Language pair</th>
<th>#Sentences</th>
<th>#unique words (L1-L2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Fr</td>
<td>2000000</td>
<td>127405-147812</td>
</tr>
<tr>
<td>En-Cr (train + tune)</td>
<td>25010</td>
<td>16294-17389</td>
</tr>
<tr>
<td>En-Cr (test)</td>
<td>284 (142 short + 142 long)</td>
<td>1168-1070 + 3562-3326</td>
</tr>
<tr>
<td>Fr-Cr</td>
<td>18354</td>
<td>13769-13725</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th></th>
<th>Short-Sentences</th>
<th>Long Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR-EN-BASIC</td>
<td>22.13</td>
<td>21.61</td>
</tr>
<tr>
<td>CR-EN-MOD</td>
<td>20.44</td>
<td>18.37</td>
</tr>
<tr>
<td>CR-EN-COMBO</td>
<td>22.74</td>
<td>21.66</td>
</tr>
<tr>
<td>CR-EN-BAC KOFF -1</td>
<td>23.69</td>
<td>22.2</td>
</tr>
<tr>
<td>CR-EN-BAC KOFF -2</td>
<td>23.69</td>
<td>22.2</td>
</tr>
<tr>
<td>CR-EN-BAC KOFF -3</td>
<td>23.58</td>
<td>21.59</td>
</tr>
<tr>
<td>CR-EN-BAC KOFF -4</td>
<td>24.9</td>
<td>21.36</td>
</tr>
<tr>
<td>LINK</td>
<td>l=1k</td>
<td>l=2k</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>DIRECT_l</td>
<td>8.86</td>
<td>11.39</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_BN</td>
<td>14.34</td>
<td>16.51</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_GU</td>
<td>13.91</td>
<td>16.15</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_KK</td>
<td>13.68</td>
<td>15.88</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_ML</td>
<td>11.22</td>
<td>13.04</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_MA</td>
<td>13.3</td>
<td>15.27</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_PU</td>
<td>15.63</td>
<td>17.62</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_TA</td>
<td>12.36</td>
<td>14.09</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_UR</td>
<td>15.34</td>
<td>17.37</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_PU_UR</td>
<td>20.53</td>
<td>21.3</td>
</tr>
</tbody>
</table>
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 \[31\]

Transformer \[8\]
Language independent NMT

- **Languages chosen:**

<table>
<thead>
<tr>
<th>Language</th>
<th>Hindi</th>
<th>Punjabi</th>
<th>Bengali</th>
<th>Gujarati</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
<th>Malayalam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>hi</td>
<td>pa</td>
<td>bn</td>
<td>gu</td>
<td>mr</td>
<td>ta</td>
<td>te</td>
<td>ml</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Tune set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46277</td>
<td>2000</td>
<td>500</td>
</tr>
</tbody>
</table>
En <-> \{Mr, Hi\} transformer output

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

<table>
<thead>
<tr>
<th></th>
<th>BLEU w/ BPE-0k</th>
<th>BLEU w/ BPE-2.5k</th>
<th>BLEU w/ BPE-5k</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Mr</td>
<td>10.79</td>
<td>14.26</td>
<td>13.48</td>
<td>10.17</td>
</tr>
<tr>
<td>Mr-En</td>
<td>19.79</td>
<td>23.82</td>
<td>24.19</td>
<td>15.87</td>
</tr>
<tr>
<td>En-Hi</td>
<td>23.77</td>
<td>29.18</td>
<td>28.92</td>
<td>26.53</td>
</tr>
<tr>
<td>Hi-En</td>
<td>24.22</td>
<td>31.22</td>
<td>30.39</td>
<td>28.15</td>
</tr>
</tbody>
</table>
En <-> {Mr, Hi} Dataset: ILCI1 + PMIndia

Dataset size (no. of sentences):

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Tune</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Hi</td>
<td>100267</td>
<td>1068</td>
<td>4273</td>
</tr>
<tr>
<td>En-Mr</td>
<td>80602</td>
<td>861</td>
<td>3445</td>
</tr>
<tr>
<td>Hi-Mr</td>
<td>92981</td>
<td>1000</td>
<td>4000</td>
</tr>
</tbody>
</table>
En <-> \{Mr, Hi\} Baselines on ILCI1+PMIndia

<table>
<thead>
<tr>
<th></th>
<th>NMT BPE-2.5k</th>
<th>NMT BPE-5k</th>
<th>NMT BPE-7.5k</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Mr</td>
<td>14.51</td>
<td>15.04</td>
<td>15.08</td>
<td>10.51</td>
</tr>
<tr>
<td>Mr-En</td>
<td>23.76</td>
<td>24.15</td>
<td>24.13</td>
<td>16.6</td>
</tr>
<tr>
<td>En-Hi</td>
<td>27.05</td>
<td>27.72</td>
<td>27.89</td>
<td>20.75</td>
</tr>
<tr>
<td>Hi-En</td>
<td>30.96</td>
<td>31.86</td>
<td>30.45</td>
<td>24.05</td>
</tr>
<tr>
<td>Hi-Mr</td>
<td>27.25</td>
<td>27.39</td>
<td>-</td>
<td>24.38</td>
</tr>
<tr>
<td>Mr-Hi</td>
<td>37.39</td>
<td>37.75</td>
<td>-</td>
<td>34.31</td>
</tr>
</tbody>
</table>
En <-> {Mr, Hi} PTI results on ILCI1+PMIndia

<table>
<thead>
<tr>
<th></th>
<th>BPE-2.5k</th>
<th>BPE-5k</th>
<th>BPE-7.5k</th>
<th>BPE-10k</th>
<th>Best Baseline BLEU</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Mr</td>
<td>14.63</td>
<td>15.97</td>
<td>15.69</td>
<td>-</td>
<td>14.51</td>
<td>+1.46</td>
</tr>
<tr>
<td>Mr-En</td>
<td>23.08</td>
<td>25.22</td>
<td>25.26</td>
<td>25.03</td>
<td>24.15</td>
<td>+1.11</td>
</tr>
<tr>
<td>En-Hi</td>
<td>25.96</td>
<td>28.79</td>
<td>29.28</td>
<td>29.23</td>
<td>27.89</td>
<td>+1.39</td>
</tr>
<tr>
<td>Hi-En</td>
<td>29.94</td>
<td>33.6</td>
<td>33.93</td>
<td>34.6</td>
<td>31.86</td>
<td>+2.74</td>
</tr>
<tr>
<td>Hi-Mr</td>
<td>-</td>
<td>27.98</td>
<td>28.49</td>
<td>-</td>
<td>27.39</td>
<td>+0.59</td>
</tr>
<tr>
<td>Mr-Hi</td>
<td>-</td>
<td>38.94</td>
<td>39.4</td>
<td>-</td>
<td>37.75</td>
<td>+1.65</td>
</tr>
</tbody>
</table>
En <-> \{Mr, Hi\} PTI + Back Translation (BT)

<table>
<thead>
<tr>
<th></th>
<th>NMT BPE-5k</th>
<th>NMT BPE-7.5k</th>
<th>Improvement over PTI model</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Mr</td>
<td>16.73 -</td>
<td></td>
<td>+0.76</td>
</tr>
<tr>
<td>Mr-En</td>
<td>-</td>
<td>26.24</td>
<td>+0.98</td>
</tr>
<tr>
<td>En-Hi</td>
<td>-</td>
<td>30.08</td>
<td>+0.8</td>
</tr>
<tr>
<td>Hi-En</td>
<td>-</td>
<td>35.03</td>
<td>+1.1</td>
</tr>
</tbody>
</table>
En <-> \{Mr, Hi\} PTI + Forward Translation (FT)

<table>
<thead>
<tr>
<th></th>
<th>NMT BPE-5k</th>
<th>NMT BPE-7.5k</th>
<th>Improvement over PTI model</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Mr</td>
<td>16.47</td>
<td>-</td>
<td>+0.5</td>
</tr>
<tr>
<td>Mr-En</td>
<td>-</td>
<td>25.9</td>
<td>+0.64</td>
</tr>
<tr>
<td>En-Hi</td>
<td>-</td>
<td>29.77</td>
<td>+0.49</td>
</tr>
<tr>
<td>Hi-En</td>
<td>-</td>
<td>34.48</td>
<td>+0.52</td>
</tr>
<tr>
<td></td>
<td>NMT BPE-5k</td>
<td>NMT BPE-7.5k</td>
<td>NMT BPE-10k</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>En-Mr</td>
<td>21.05</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mr-En</td>
<td>28.76</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>En-Hi</td>
<td>-</td>
<td>-</td>
<td>34.32</td>
</tr>
<tr>
<td>Hi-En</td>
<td>-</td>
<td>-</td>
<td>38.65</td>
</tr>
<tr>
<td>Hi-Mr</td>
<td>-</td>
<td>29.09</td>
<td>-</td>
</tr>
<tr>
<td>Mr-Hi</td>
<td>-</td>
<td>36.98</td>
<td>-</td>
</tr>
</tbody>
</table>
## En-Hi-Mr NMT with embellishments (consolidated)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Hi-En</th>
<th>En-Hi</th>
<th>Hi-Mr</th>
<th>Mr-Hi</th>
<th>En-Mr</th>
<th>Mr-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>C: Agnostic Training</td>
<td>41.9</td>
<td>37.95</td>
<td>29.1</td>
<td>37.08</td>
<td><strong>25.34</strong></td>
<td>32.4</td>
</tr>
<tr>
<td>D: PTI(Phrase Table Injection)</td>
<td>38.65</td>
<td>34.32</td>
<td>29.09</td>
<td>36.98</td>
<td>21.05</td>
<td>28.76</td>
</tr>
<tr>
<td>E: D + Enhancement</td>
<td>42.15</td>
<td>36.78</td>
<td><strong>29.18</strong></td>
<td>37.23</td>
<td><strong>21.91</strong></td>
<td>29.57</td>
</tr>
<tr>
<td>F: BERT augmented NMT</td>
<td>29.74</td>
<td>25.89</td>
<td>28.64</td>
<td>34.21</td>
<td>14.98</td>
<td>18.37</td>
</tr>
<tr>
<td>G: BPE+word(pretrained BPE embeddings)</td>
<td>49.65</td>
<td>43.05</td>
<td>25.40</td>
<td>31.45</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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 माहीत 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 धार्मिक प्रेरणारे महानायक होते .
OP: आज आपल्या देशात लोहपुरुष सरदार वल्लभभाई पटेल यांची जयंती , एक देश महून एकत्र आणत आहेत .
GT: हा दिवस भारतीय लोहपुरुष सरदार वल्लभभाई पटेल या जयंतीनिमित आम्हाला राष्ट्र महून जोडण्याचे एकत्रीकरण
Bing: या दिवशी सरदार वल्लभभाई पटेल या लोहपुरुषाची जयंती आहे . (incomplete)
Hi: यदि श्वास प्रणालिका में सूजन आ जाये तब भी रक्त खुंख के रास्ते बाहर आने लगता है।

Reference Mr: जर श्वासनिलिकेत सूज आली तरीही रक्त तोंडावाटे बाहर येऊ लागते।

Model Op: जर श्वासनिलिकेत सूज आली तरीही रक्त तोंडावाटे बाहर 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 भूख रहता है और मस्तिष्क को भूख के संकेत देता है।

Reference Mr: जेक्हा हे भाग वेगाना कमी होतात, तेक्हा पोट थोडेसे भूके राहतात।

Model Op: जेक्हा हा भाग तीव्रतेने कमी होत असतो तेक्हा पोट थोडे उपाशी राहतो आणि मस्तिष्काला भूकेचा संकेत देतो।

Google: जेक्हा हे भाग झपाट्याने कमी होतात तेक्हा पोट किचित भूक राहते आणि मेंदुला 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 घरमैती काम के ले जाने अवर्णनीय निर्धारित करून पहिले या सुरू के ले जाने

Google: आरोग्यास परवानगी मिळाल्यास रजतील काम नियमितपणे करावीत. व्यावसायिक काम श्रेणीरित्या पहिले सुरू करवे लागेल.

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

Model Op: रोज सकाळी किंवा संध्याकाळी आरामपट चालणे जेवढे सहन के ले जाऊ शकते.

Google: जतके सहन के ले जाऊ शकते तितके दररोज सकाळी किंवा संध्याकाळी आरामपट चालणे.
Examples from Programming Domain

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

Model Op: दूसर्या बाजूला एक अपायकारक अनावश्यक वापर केला जातो जेव्हा तुम्हाला लगेच कठिन नाही.

Google: दूसरीकडे जेव्हा आपल्याला आणाऱ्यास माहिती नसते तेच्या पाबळा पठवाव सामान्यत: वापरला जातो.

Hi: अब हम यह अंत से शुरू कर रहे हैं और केवल पहला कारक रख रहे हैं 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 अनावश्यक पाहिले , एक नवीन प्रकारचा स्पष्ट आहे.

Google: आता आपण या टॉकपासून सुरूवात करून ती आहोत आणि फक्त पहिला घटक ठेवणे म्हणजे आपण, नंतर या उदाहरणात जे पाहिले ते एक नवीन प्रकारचे लूप आहें.
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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Set</th>
<th>Disfluent Sentences</th>
<th>Fluent Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Parallel</td>
<td>Train</td>
<td>55,482</td>
<td>55,482</td>
</tr>
<tr>
<td>Parallel</td>
<td>Dev</td>
<td>11,889</td>
<td>11,889</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>11,889</td>
<td>11,889</td>
</tr>
<tr>
<td>Model</td>
<td>Validation</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence to Sequence (Bi-LSTM)</td>
<td>87.23</td>
<td>88.08</td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td>89.27</td>
<td>90.08</td>
<td></td>
</tr>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise Induction (Transformer)</td>
<td>65.17</td>
<td>53.78</td>
<td></td>
</tr>
<tr>
<td>Style Transfer (Bi-LSTM)</td>
<td>61.26</td>
<td>62.77</td>
<td></td>
</tr>
<tr>
<td>Style Transfer (Transformer)</td>
<td>78.72</td>
<td>79.39</td>
<td></td>
</tr>
<tr>
<td>Semi-Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Style Transfer (Transformer)</td>
<td>84.1</td>
<td>85.28</td>
<td></td>
</tr>
</tbody>
</table>

Semi-Supervised:

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

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

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 Kanojia
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
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 $W^{s \rightarrow t}$. 

* Image source: www.mikelartetxe.com
Cross-lingual embeddings: Approaches

- Mapping based
  - Vecmap
  - MUSE
  - GeoMM
  - RCSLS

- Joint loss based
  - Joint - Replace
  - Joint Matrix Factorization

- Pseudo multi-lingual corpora based
  - Random Translation Replacement
  - Multilingual Cluster
Cross-lingual embeddings: Mapping based

- Task is to learn $W_X$ and $W_Y$ (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.

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
  - $W_X = US^{1/2}$, $W_Y = VS^{1/2}$

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

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

- 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)

Monolingual data of L1 → MT system (L2->L1) → Translated sentences

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

Iterative Back-Translation

Train $\text{MT}_{L2\rightarrow L1}$ using $D$

Generate synthetic data $(SD)$ for $\text{MT}_{L1\rightarrow L2}$ using $\text{MT}_{L2}$.

$D = D \cup SD$

Generate synthetic data $(SD)$ for $\text{MT}_{L2\rightarrow L1}$ using $\text{MT}_{L1}$.

$D = D \cup SD$

Train $\text{MT}_{L1\rightarrow L2}$ using $D$

$D = \text{Parallel corpus}$

$SD = \text{Synthetic data}$
Iterative Back-Translation

<table>
<thead>
<tr>
<th>Setting</th>
<th>French–English 100K</th>
<th>English–French 100K</th>
<th>Farsi–English 100K</th>
<th>English–Farsi 100K</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT baseline</td>
<td>16.7</td>
<td>24.7</td>
<td>21.7</td>
<td>16.4</td>
</tr>
<tr>
<td>back-translation</td>
<td>22.1</td>
<td>27.8</td>
<td>22.1</td>
<td>16.7</td>
</tr>
<tr>
<td>back-translation iterative+1</td>
<td>22.5</td>
<td>-</td>
<td>22.7</td>
<td>17.1</td>
</tr>
<tr>
<td>back-translation iterative+2</td>
<td>22.6</td>
<td>-</td>
<td>22.6</td>
<td>17.2</td>
</tr>
</tbody>
</table>

- Beneficial for Low resource languages also

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


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

Training Procedure

For $n$ iterations

$\text{DAE}_{\text{src}} \rightarrow \text{DAE}_{\text{trg}} \rightarrow \text{BTS}_{\text{src}} \rightarrow \text{BTS}_{\text{trg}}$

$\text{DAE}_{\text{src}}$: Denoising of source sentences; $\text{DAE}_{\text{trg}}$: Denoising of target sentences;
$\text{BTS}_{\text{src}}$: Back-translation with shuffled source sentences; $\text{BTS}_{\text{trg}}$: Back-translation with shuffled target sentences;
$n$: total number of iteration till it reaches stopping criterion.
U-NMT: Denoising of source sentences

U-NMT: Denoising of target sentences

Input trg sentence → Noisy adding algorithm → Noisy trg sentence → Trainable unit

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

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

Input real trg sentence → Noise adding algorithm → le disque comprendra aussi deux chansons en italien → Output in src language (synthetic sentence)

the disc will also include two songs in Italian

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

The disc will also include two songs in Italian.

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

Comparison between two approaches

- 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

Unsupervised dictionary induction

सुरक्षित (surakshit)

घर पर सुरक्षित रहें (ghar par surakshit rahen)

Generation of word-translated sentence
Training with word-by-word translation

Unsupervised dictionary induction

Generation of synthetic parallel corpus

Hindi monolingual corpus → Unsupervised dictionary induction → Synthetic Gujarati translations

घर पर सुरक्षित रहें (ghar par surakshit rahen)

घर चालू सलामत रहेवृं (Ghar Cālu salāmata rahēvṛum)
Training with word-by-word translation

Can we use this synthetic parallel corpus to train a NMT model?
Training with word-by-word translation

घर पर सुरक्षित रहें (ghar par surakshit rahen) →  Unsupervised dictionary induction → घर चालू सलामत रहेवूं (Ghar ālu salāmata rahēvuṁ)

घर पर सुरक्षित रहें (ghar par surakshit rahen) → UNMT → घरे सलामत रहेवूं (Gharē salāmata rahēvuṁ)
## Effect of DAE and BT

<table>
<thead>
<tr>
<th>Author</th>
<th>Approach</th>
<th>Fr $\rightarrow$ En</th>
<th>En $\rightarrow$ Fr</th>
<th>De $\rightarrow$ En</th>
<th>En $\rightarrow$ De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artette et al. (tested on WMT14)</td>
<td>Emb. nearest neighbour</td>
<td>9.98</td>
<td>6.25</td>
<td>7.07</td>
<td>4.39</td>
</tr>
<tr>
<td></td>
<td>Denoising</td>
<td>7.28</td>
<td>5.33</td>
<td>3.64</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>Denoising + Back-translation</td>
<td>15.56</td>
<td>15.13</td>
<td>10.21</td>
<td>6.55</td>
</tr>
<tr>
<td>Lample et al. (tested on WMT14 en-fr and WMT16 en-de)</td>
<td>Emb. nearest neighbour</td>
<td>10.09</td>
<td>6.28</td>
<td>10.77</td>
<td>7.06</td>
</tr>
<tr>
<td></td>
<td>Word2word pretraining</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ Denoising</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ Back-translation</td>
<td>15.31</td>
<td>15.05</td>
<td>13.33</td>
<td>9.64</td>
</tr>
</tbody>
</table>


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

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

Shared Encoder

Shared Decoder

L1 Encoder

L2 Encoder

L1 Decoder

L2 Decoder

Shared path

Path for L1

Path for L2

Shared path
Language specific Encoder-Decoder

- L1 Encoder
- L1 Decoder
- L2 Encoder
- L2 Decoder

Path for L1
Path for L2
Shared path
Language specific Encoder-Decoder

How to share Latent space?
Parameter sharing

Layer 1 of lang1
Layer 2 of lang1
Layer 3 of lang1
Layer 4 of lang1

Layer 1 of lang2
Layer 2 of lang2
Layer 3 of lang2
Layer 4 of lang2

output

chocolate in the box

बॉक्स में चॉकलेट
Parameter sharing

Layer 1 of lang1

Layer 2 of lang1

Layer 3 of lang1

Layer 1 of lang2

Layer 2 of lang2

Layer 3 of lang2

Shared layer 4

output

chocolate in the box

बॉक्स में चॉकलेट
Parameter sharing

Layer 1 of lang1

Layer 1 of lang2

Layer 2 of lang1

Layer 2 of lang2

Shared layer 3

Shared layer 4

output

chocolate in the box

बॉक्स में चॉकलेट
Parameter sharing

Layer 1 of lang1

Layer 1 of lang2

Shared layer 2

Shared layer 3

Shared layer 4

output

chocolate in the box

बॉक्स में चॉकलेट
Parameter sharing

Shared layer 1

Shared layer 2

Shared layer 3

Shared layer 4

output

chocolate in the box
Architecture with weight-sharing layers

Layer 1 of lang 1
Layer 3 of lang 1
Layer 2 of lang 1
Layer 1 of lang 2

Layer 2 of lang 1
Layer 3 of lang 2
Layer 4 of lang 1
Layer 4 of lang 2

Shared layer 1

Layer 3 of lang 1
Layer 3 of lang 2

Shared layer 4

Layer 2 of lang 1
Layer 2 of lang 2
Layer 1 of lang 1
Layer 1 of lang 2

Latent space

chocolate in the box

बॉक्स में चॉकलेट
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.

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.

Results

<table>
<thead>
<tr>
<th></th>
<th>en-de</th>
<th>de-en</th>
<th>en-fr</th>
<th>fr-en</th>
<th>zh-en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>24.07</td>
<td>26.99</td>
<td>30.50</td>
<td>30.21</td>
<td>40.02</td>
</tr>
<tr>
<td>Word-by-word</td>
<td>5.85</td>
<td>9.34</td>
<td>3.60</td>
<td>6.80</td>
<td>5.09</td>
</tr>
<tr>
<td>Lample et al. (2017)</td>
<td>9.64</td>
<td>13.33</td>
<td>15.05</td>
<td>14.31</td>
<td>-</td>
</tr>
<tr>
<td>The proposed approach</td>
<td><strong>10.86</strong></td>
<td><strong>14.62</strong></td>
<td><strong>16.97</strong></td>
<td><strong>15.58</strong></td>
<td><strong>14.52</strong></td>
</tr>
</tbody>
</table>

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**


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
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

<table>
<thead>
<tr>
<th>W</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update</td>
<td>Update</td>
</tr>
</tbody>
</table>
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.

\[ L = L_{\text{cycle}}(E) + L_{\text{cycle}}(F) + L_{\text{lm}}(E) + L_{\text{lm}}(F) \]
Iterative refinement

- Generate a synthetic parallel corpus by translating the monolingual corpus with the initial system $L_1 \rightarrow L_2$, and train and tune SMT system $L_2 \rightarrow L_1$.
  - 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.

\[
\text{score}(f|e) = \prod_i \max \left( \epsilon, \max_j \text{sim}(\tilde{f}_i, \tilde{e}_j) \right)
\]
## Results

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

UMT Approaches

1. Unsupervised NMT
2. GAN for UNMT
3. Unsupervised SMT
4. Hybrid UMT
We can combine UNMT and USMT in two ways.
  ○ USMT followed by UNMT.
  ○ UNMT followed by USMT.

List of papers


USMT followed by UNMT Vs. UNMT followed by USMT

- **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.

---

<table>
<thead>
<tr>
<th>WMT 14/16</th>
<th>En→Fr</th>
<th>Fr→En</th>
<th>En→De</th>
<th>De→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT + PBSMT</td>
<td>27.1</td>
<td>26.3</td>
<td>17.5</td>
<td>22.1</td>
</tr>
<tr>
<td>PBSMT + NMT</td>
<td>27.6</td>
<td>27.7</td>
<td>20.2</td>
<td>25.2</td>
</tr>
</tbody>
</table>

---

Pre-training approaches for Unsupervised NMT
XLM, CMLM, MASS, BART, mBART
XLM

Cross-lingual Language Modelling Pre-Training

Typical Deep Learning Module

Input Symbol
(Characters, Words, Phrases, Sentences...)

Lookup Table

Encoder

Decoder

Output

Embedding Space
(Randomly Initialized Lookup Table)

bank
rain
humidity
finance
weather
money
Typical Deep Learning Module

Output

Decoder

Encoder

Lookup Table
(Initialized Pre-trained embeddings)

Input Symbol
(Characters, Words, Phrases, Sentences…)

Embedding Space
(Initialized with Pre-Trained Embeddings)

money
bank
finance
weather
rain
humidity

(Bengio et.al 2003, Collobert and Weston 2011, Mikolov et.al 2013)
Typical Deep Learning Module

Input Symbol
(Characters, Words, Phrases, Sentences…)

Lookup Table
(Initialized Pre-trained embeddings)

Encoder

Decoder

Output

I went to the market
I went to the grocery store
India won the match
India defeated Australia

Pre-train Encoder?
General Framework

L1 Monolingual Corpus

L2 Monolingual Corpus

Language Model Pre-Training

Unsupervised NMT Fine Tuning

Pre-Training

Fine-Tuning
XLM Pre-Training
XLM Fine Tuning

- Perform fine-tuning using
  - Iterative back-translation
  - Denoising auto-encoding
- Alternate between the two objectives
- **Denoising auto-encoding** helps in better training of the decoder
## XLM: Results

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

- 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

\[ P(y_{1}^{m} | x_{1}^{l}) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i, |j, l, m) P(y_{j} | x_{i}) \]

\( \epsilon \) = 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

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

\( \epsilon \) = probability that the translation of \( x_1^l \) consists of \( m \) tokens
\( a(i, \mid j, l, m) \) = probability that \( i^{th} \) source token is aligned to \( j^{th} \) target token

- The loss function becomes

\[ L_{cmlm} = -\log (\epsilon) - \sum_{j=1}^{m} \log \left( \sum_{i=0}^{l} a(i, \mid j, l, m) P(y_j \mid x_i) \right) \]
Cross-lingual Masked Language Modelling

Modeling

- The loss function becomes

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

- The gradient becomes:

\[ \nabla L = \sum_{j=1}^{m} \frac{a(i | j, l, m) P(y_j | x_i)}{\sum_{i=0}^{l} a(i | j, l, m) P(y_j | x_i)} \nabla \log P(y_j | x_i) \]
Cross-lingual Masked Language Modelling

Modeling

- The gradient becomes:

\[ \nabla L = \sum_{j=1}^{m} \frac{a(i, j, l, m) \cdot P(y_j | x_i)}{\sum_{l=0}^{l} a(i, j, l, m) \cdot P(y_j | x_i)} \cdot \nabla \log P(y_j | x_i) \]

- \( a(i, |j, l, m) \) are approximated using cross-lingual BPE embedding

- \( P(y_j | x_i) \) 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

<table>
<thead>
<tr>
<th>Method</th>
<th>fr2en</th>
<th>en2fr</th>
<th>de2en</th>
<th>en2de</th>
<th>ro2en</th>
<th>en2ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Artetxe et al., 2017)</td>
<td>15.6</td>
<td>15.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Lample et al., 2017)</td>
<td>14.3</td>
<td>15.1</td>
<td>13.3</td>
<td>9.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Artetxe et al., 2018b)</td>
<td>25.9</td>
<td>26.2</td>
<td>23.1</td>
<td>18.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Lample et al., 2018)</td>
<td>27.7</td>
<td>28.1</td>
<td>25.2</td>
<td>20.2</td>
<td>23.9</td>
<td>25.1</td>
</tr>
<tr>
<td>(Ren et al., 2019)</td>
<td>28.9</td>
<td>29.5</td>
<td>26.3</td>
<td>21.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Lample and Conneau, 2019)</td>
<td>33.3</td>
<td>33.4</td>
<td>34.3</td>
<td>26.4</td>
<td>31.8</td>
<td>33.3</td>
</tr>
<tr>
<td>Iter 1 (CMLM)</td>
<td>34.8</td>
<td>34.9</td>
<td>35.5</td>
<td>27.9</td>
<td>33.6</td>
<td>34.7</td>
</tr>
<tr>
<td>Iter 2 (CMLM)</td>
<td><strong>34.9</strong></td>
<td><strong>35.4</strong></td>
<td><strong>35.6</strong></td>
<td><strong>27.7</strong></td>
<td><strong>34.1</strong></td>
<td><strong>34.9</strong></td>
</tr>
</tbody>
</table>
CMLM
Cross-lingual Masked Language Modelling

Ablation Study
CMLM: Ablation Study

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

<table>
<thead>
<tr>
<th></th>
<th>fr2en</th>
<th>en2fr</th>
<th>de2en</th>
<th>en2de</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CMLM + MLM</strong></td>
<td>34.8</td>
<td>34.9</td>
<td>35.5</td>
<td>27.9</td>
</tr>
<tr>
<td><strong>CMLM</strong></td>
<td>34.1</td>
<td>34.3</td>
<td>35.1</td>
<td>27.2</td>
</tr>
<tr>
<td><strong>- translation prediction</strong></td>
<td>33.7</td>
<td>33.9</td>
<td>34.8</td>
<td>26.6</td>
</tr>
<tr>
<td><strong>- - n-gram mask</strong></td>
<td>33.3</td>
<td>33.4</td>
<td>34.3</td>
<td>26.4</td>
</tr>
</tbody>
</table>

**CMLM + MLM** means we use $L_{\text{pre}}$ as the pre-training loss;
**CMLM** means we only use $L_{\text{cmlm}}$ 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

Diagram showing the Encoder and Decoder + Attention layers with input tokens 'I', 'went', 'to', and 'ket' and output tokens 'the' and 'mar@@'.
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)

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

*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, $X = \text{I went to the market yesterday night}$
  - Let $\text{to the market yesterday}$ be the text segment selected for masking
  - Default value is 50% of the input sentence
  - However, not all tokens $\text{to the market yesterday}$ 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 $k = 80\%$, $l = 10\%$
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

Select randomly 50% of the consecutive tokens for masking

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

Output to be generated
## MASS: Word Selection Hyper-parameters

<table>
<thead>
<tr>
<th>Configuration</th>
<th>%age Masked</th>
<th>%age Retained</th>
<th>%age Randomly replaced</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>-</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>-</td>
<td>90</td>
</tr>
</tbody>
</table>
MASS (Song et al. 2019): Word Selection Hyper-parameters
# MASS: Word Selection Hyper-parameters

<table>
<thead>
<tr>
<th>Configuration</th>
<th>%age Masked</th>
<th>%age Retained</th>
<th>%age Randomly replaced</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>60</td>
<td>20</td>
<td>Auto-encoder</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>Auto-encoder</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>20</td>
<td>20</td>
<td>Auto-encoder</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>10</td>
<td>10</td>
<td>Recommended</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>5</td>
<td>5</td>
<td>Recommended</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>20</td>
<td>60</td>
<td>Unable to generate translations. But perplexity is low (Better for other tasks?)</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>-</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>-</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>
MASS (Song et.al 2019): Role of Masking Tokens

- Consider the input sentence, \( X = I \text{ went to the market yesterday night} \)
- Let \textit{to the market yesterday} be the text segment selected for masking
- The input to the encoder is \( I \text{ went } _ _ _ _ \text{ night} \)
- The input to the decoder (previous token) is \textit{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

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Method</th>
<th>BLEU</th>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>36.9</td>
<td>Feed</td>
<td>35.3</td>
<td>MASS</td>
<td>37.5</td>
</tr>
</tbody>
</table>

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

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

Original document

My name is John. I go to school daily.

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

mBART (Liu et al 2020)

- Pre-training using BART objective on multiple languages

<table>
<thead>
<tr>
<th>Model</th>
<th>Similar Pairs</th>
<th>Dissimilar Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-De</td>
<td>En-Ro</td>
</tr>
<tr>
<td>Random</td>
<td>21.0</td>
<td>17.2</td>
</tr>
<tr>
<td>XLM (2019)</td>
<td>34.3</td>
<td>26.4</td>
</tr>
<tr>
<td>MASS (2019)</td>
<td>35.2</td>
<td>28.3</td>
</tr>
<tr>
<td>mBART</td>
<td>34.0</td>
<td>29.8</td>
</tr>
</tbody>
</table>

- 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

When Unsupervised NMT does not work?


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

<table>
<thead>
<tr>
<th>Domain (en)</th>
<th>Domain (de/ru)</th>
<th>BLEU [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de-en</td>
<td>en-de</td>
</tr>
<tr>
<td>Newswire</td>
<td>23.3</td>
<td>19.9</td>
</tr>
<tr>
<td>Politics</td>
<td>11.5</td>
<td>12.2</td>
</tr>
<tr>
<td>Random</td>
<td>18.4</td>
<td>16.4</td>
</tr>
</tbody>
</table>

- Different distributions of the topics

Initialization

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

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

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

Monolingual Word Representations
(capture syntactic and semantic similarities between words)

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

(Source: Khapra and Chandar, 2016)
Why is the Quality questioned?

Encode-Decode paradigm used for MT
Good Quality Cross-lingual Embeddings?

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

Encode-Decode paradigm used for MT
**Quantitative Quality**

<table>
<thead>
<tr>
<th>Source - Target</th>
<th>GeoMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>En - Es</td>
<td>81.4</td>
</tr>
<tr>
<td>Es - En</td>
<td>85.5</td>
</tr>
<tr>
<td>En - Fr</td>
<td>82.1</td>
</tr>
<tr>
<td>Fr - En</td>
<td>84.1</td>
</tr>
<tr>
<td>En - De</td>
<td>74.7</td>
</tr>
<tr>
<td>De - En</td>
<td>76.7</td>
</tr>
<tr>
<td>En - Hi</td>
<td>41.5</td>
</tr>
<tr>
<td>Hi - En</td>
<td>54.8</td>
</tr>
<tr>
<td>En - Ta</td>
<td>31.9</td>
</tr>
<tr>
<td>Ta - En</td>
<td>38.7</td>
</tr>
<tr>
<td>En - Bn</td>
<td>36.7</td>
</tr>
<tr>
<td>Bn - En</td>
<td>42.7</td>
</tr>
</tbody>
</table>

Very low Precision@1 for Indic languages compared to the European language counterpart.
## Unsupervised NMT [Lample et al 2018]

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

<table>
<thead>
<tr>
<th></th>
<th>Multi30k-Task1</th>
<th></th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en-fr</td>
<td>fr-en</td>
<td>de-en</td>
</tr>
<tr>
<td>Supervised</td>
<td>56.83</td>
<td>50.77</td>
<td>38.38</td>
</tr>
<tr>
<td>word-by-word</td>
<td>8.54</td>
<td>16.77</td>
<td>15.72</td>
</tr>
<tr>
<td>word reordering</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Unsupervised NMT [Lample et al 2018]

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

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>En → Fr</td>
<td>6.28</td>
</tr>
<tr>
<td>Fr → En</td>
<td>10.09</td>
</tr>
<tr>
<td>En → De</td>
<td>7.06</td>
</tr>
<tr>
<td>De → En</td>
<td>10.77</td>
</tr>
<tr>
<td>En → Hi</td>
<td>1.2</td>
</tr>
<tr>
<td>Hi → En</td>
<td>2.1</td>
</tr>
</tbody>
</table>

**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

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

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 (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
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.
4. Better cross-lingual language model pretraining for resource-scarce languages, disimilar languages, and dissimilar domains.
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

References


References

References

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

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

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Comments</th>
<th>Google Translate [30 Nov, 2020]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ನಾನು ಹೇಳುವುದನುಂ ಸರಿಯಾಗಿ ಕೇಲಿಸಿಕೋ</td>
<td>Literary Language</td>
<td>Listen to me correctly</td>
</tr>
<tr>
<td>ನಾನು ಹೇಳೋದನ್ನು ಸರಿಯಾಗಿ ಕೇಲಸು</td>
<td>Spoken Language</td>
<td>I am Sergio Katsko of Noodon</td>
</tr>
<tr>
<td>ಎಲ್ಲೆಗೆ ಮಹತ್ವದಲ್ಲಿ ತಿನಿಗೆ UTa mADikoMDu hogu</td>
<td>Literary Language</td>
<td>Go Have lunch (Go after having lunch)</td>
</tr>
<tr>
<td>ಎಲ್ಲೆಗೆ ಮಹತ್ವದಲ್ಲಿ ತಿನಿಗೆ UTa mADkoMDu hogu</td>
<td>Spoken Language</td>
<td>Modify the meal</td>
</tr>
</tbody>
</table>

Phenomenon similar to Schwa Deletion in Literary language and Spoken language
## Why Indic Languages?

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Comments</th>
<th>Google Translate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ವಂಚಕಾಸುರರನ್ನು ಒದ್ದಿರುವ ದೇನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ?</td>
<td>Maximum Sandhi transformation</td>
<td>Do you know anyone who has cheated?</td>
</tr>
<tr>
<td>ಸರ್ಕಾರ ವಂಚಕಾಸುರರನ್ನು ಒದ್ದಿರುವ ದೇನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ?</td>
<td>No Sandhi transformation</td>
<td>Do you know who became the one who drove out the crafty demons?</td>
</tr>
<tr>
<td>ಸರ್ಕಾರ ವಂಚಕಾಸುರರನ್ನು ಒದ್ದಿರುವ ದೇನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ?</td>
<td>Normal Usage</td>
<td>Do you know who is the one who kicked the crooks?</td>
</tr>
</tbody>
</table>

Crooked demons one who kicked them away who is you know
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.

\[
\text{score}(\bar{f} | e) = \prod_i \max \left( \varepsilon, \max_j \text{sim}(\bar{f}_i, \bar{e}_j) \right)
\]

\[
\text{sim}(f, e) = 1 - \frac{\text{lev}(f, e)}{\max(\text{len}(f), \text{len}(e))}
\]
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

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

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.

A B C . D E .