Machine Translation

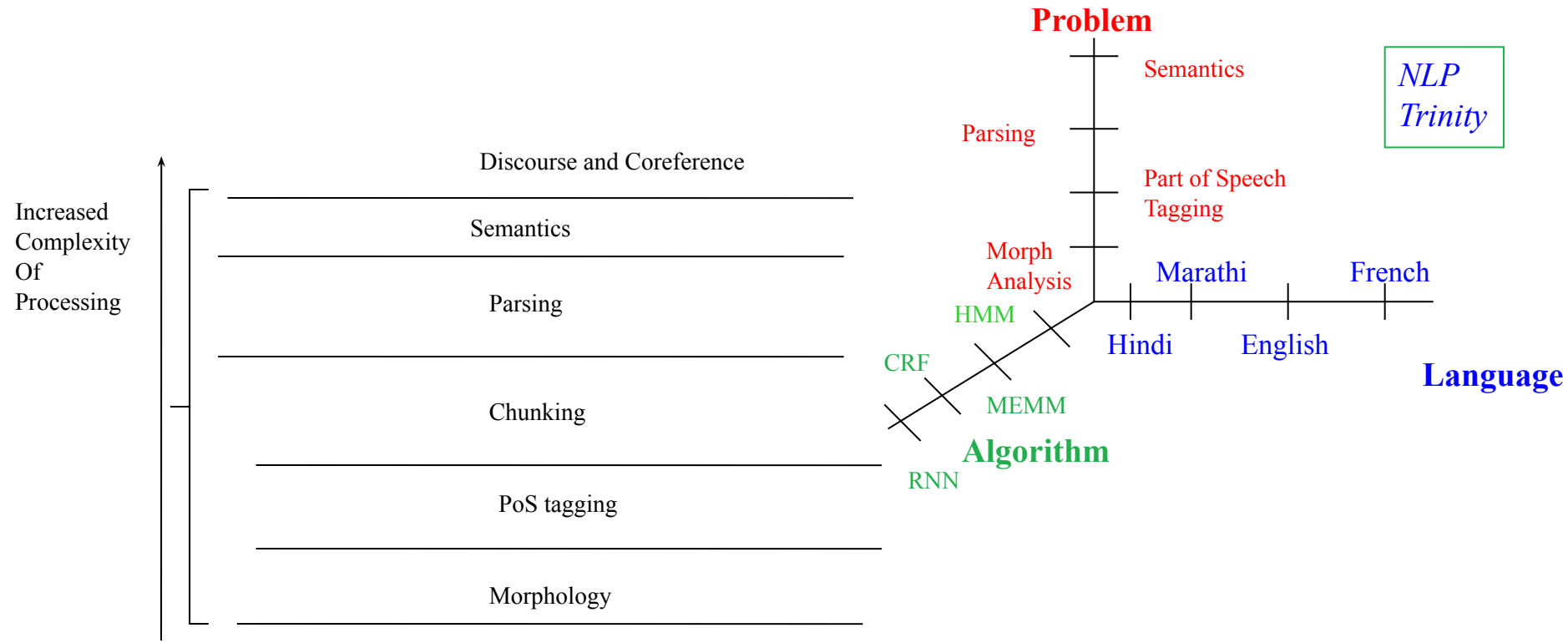
Asif Ekbal
AI-NLP-ML Group
Department of Computer Science and Engineering
IIT Patna, India
Email: asif@iitp.ac.in, asif.ekbal@gmail.com

Outline

- Introduction and Background to Machine Translation
- MT approaches
- MT in Low-resources languages
  - Transfer Learning, Domain Adaptation, Multilinguality, Unsupervision
- MT in Indian languages
  - Rule-based, EBMT
  - SMT, NMT
  - Hybrid
- Research on MT at IIT Patna
- Takeaways

Machine Translation

(One of the hardest, but fascinating & useful problem of NLP)


Increased
Complexity
Of
Processing
Discourse and Coreference
Semantics
Parsing
Chunking
PoS tagging
Morphology
Problem
Semantics
Parsing
Part of Speech
Tagging
Morph
Analysis
NLP
Trinity
Marathi
French
Hindi
English
Language
HMM
CRF
MEMM
Algorithm
RNN

Machine Translation: *From Wikipedia*

The origins of machine translation can be traced back to the work of Al-Kindi, a 9th-century Arabic cryptographer who developed techniques for systemic language translation, including cryptanalysis, frequency analysis, and probability and statistics, which are used in modern machine translation.[8] The idea of machine translation later appeared in the 17th century. In 1629, René Descartes proposed a universal language, with equivalent ideas in different tongues sharing one symbol.[9]

In the mid-1930s the first patents for "**translating machines**" were applied for by Georges Artsrouni, for an automatic bilingual dictionary using paper tape. Russian Peter Troyanskii submitted a more detailed proposal[10][11] that included both the bilingual dictionary and a method for dealing with grammatical roles between languages, based on the grammatical system of Esperanto. This system was separated into three stages: stage one consisted of a native-speaking editor in the source language to organize the words into their logical forms and to exercise the syntactic functions; stage two required the machine to "translate" these forms into the target language; and stage three required a native-speaking editor in the target language to normalize this output. Troyanskii's proposal remained unknown until the late 1950s, by which time computers were well-known and utilized

History of MT: Pessimism

- 1959/1960: Bar-Hillel “Report on the state of MT in US and GB”
  - Argued FAHQUT (fully automatic high quality translation of unrestricted text) too hard (semantic ambiguity, etc)
  - Should work on semi-automatic instead of automatic
  - His argument
    Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.
  - Only human knowledge let’s us know that ‘playpens’ are bigger than boxes, but ‘writing pens’ are smaller
  - His claim: we would have to encode all of human knowledge

History of MT: Pessimism

- The ALPAC (*Automatic Language Processing Advisory Committee*) report
  - Headed by John R. Pierce of Bell Labs
  - Conclusions:
    - Supply of human translators exceeds demand
    - All the Soviet literature is already being translated
    - **MT has been a failure:** all current MT works had to be post-edited
    - Sponsored evaluations which showed that intelligibility and informativeness were worse than human translations
  - Consequences:
    - MT research suffered
      - Funding loss
      - Number of research labs declined
      - Association for Machine Translation and Computational Linguistics dropped MT from its name

Multilinguality: Indian situation

- **Major streams**
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Some languages are ranked within 20 in the world in terms of the populations speaking them
  - Hindi : 4th (~350 milions)
  - Bangla: 5^{h} (~230 millions)
  - Marathi: 10th (~84 millions)

Background: Indian Context

- India is a multilingual country with great linguistic and cultural diversities
- 22 official languages mentioned in the Indian constitution
- However, Census of India in 2001 reported-
  - **122 major languages**
  - **1,599 other regional languages**
  - **13 scripts**
  - **30 languages** are spoken by more than **one million native speakers**
  - **122** are spoken by more than **10,000 people**
- **20%** understand English
- **80%** cannot understand

Background

- Phenomenal growth in the number of internet users, social media
  Facebook, Twitter, Blogs, Review Sites etc.
- Increasing tendency of using Indian language contents for exchanging information
- **Digital divide** cannot be tackled unless citizens are given flexibility in **communicating in their own languages**

Machine Translation can play an important role towards creating this digital society

- **Manual Translation (MT) is slow, tedious and impracticable**
- **Development of Machine Translation systems are fast, takes less time and saves money**

History of MT

- 17th century: Leibniz and Descartes put forward proposals for codes which would relate words between languages
- 1947-48: idea of dictionary-based direct translation
- 1949: Warren Weaver's Memorandum on Translation
- 1954: Public demonstration of Georgetown-IBM experiment
- 1966: ALPAC (Automatic Language Processing Advisory Committee) report
- 1970s: Rule-based systems used by companies: *Systran, Logos*
- 1990s: Statistical MT and IBM models
- Mid 2000s: Phrase-based MT (Google and Moses)
- Around 2010: Commercial viability
- Since mid 2010s: Neural Network models for MT
- 2016: Neural Machine Translation (NMT), the new state-of-the-art MT techniques

Challenges in MT (1/2)

- **Availability of data**: Development of Fully Automatic High Quality Translation (FAHQT) systems requires huge amount of parallel corpora
- **Domain adaptation**: MT systems designed for one domain might fail to translate data in other domains
- **Limited context**: Current MT systems translate each sentence independently without taking the context of other sentences which might result in non-coherent translations
- **Translation of rare words**: Current MT systems most of the time fail to translate words belonging to specific categories, such as named entities
- **Noisy data**: MT systems fail when the training data contains noise

Challenges in MT (2/2)

- **Lexical divergences:** There is a great divergence among the languages, for example
  - Word order: SVO (English), SOV (Hindi), VSO, OSV
  - Free word order (Sanskrit) vs rigid word order (English)
  - Inflectional systems: Infixing (Arabic), Agglutinative (Telugu), Fusional (Sanskrit)

- **Semantic ambiguity**: A word can have multiple meanings (polysemy) and many words can have same meaning (synonymy). Translating these types of words are always difficult
- **Morphological richness**
  - Identifying basic units of words
  - Fertility of languages

MT in different domains

- **Government**
  - Military and Defence (Systran, SDL)
  - Finance (Lingua Custodia)
  - Judicial (Systran, SDL, HEMAT)
- **Software and Technology**
  - E-commerce (Systran, Google Translate, Bing Translator)
  - E-learning (Google Translate, Bing Translator)
- **Healthcare** (Canopy Speak)
- **Social**
  - Travel (Google Translate, Bing Translator)
  - Entertainment (Google Translate, Bing Translator, Amazon Translate)
  - Social media (Google Translate, Facebook Translate)

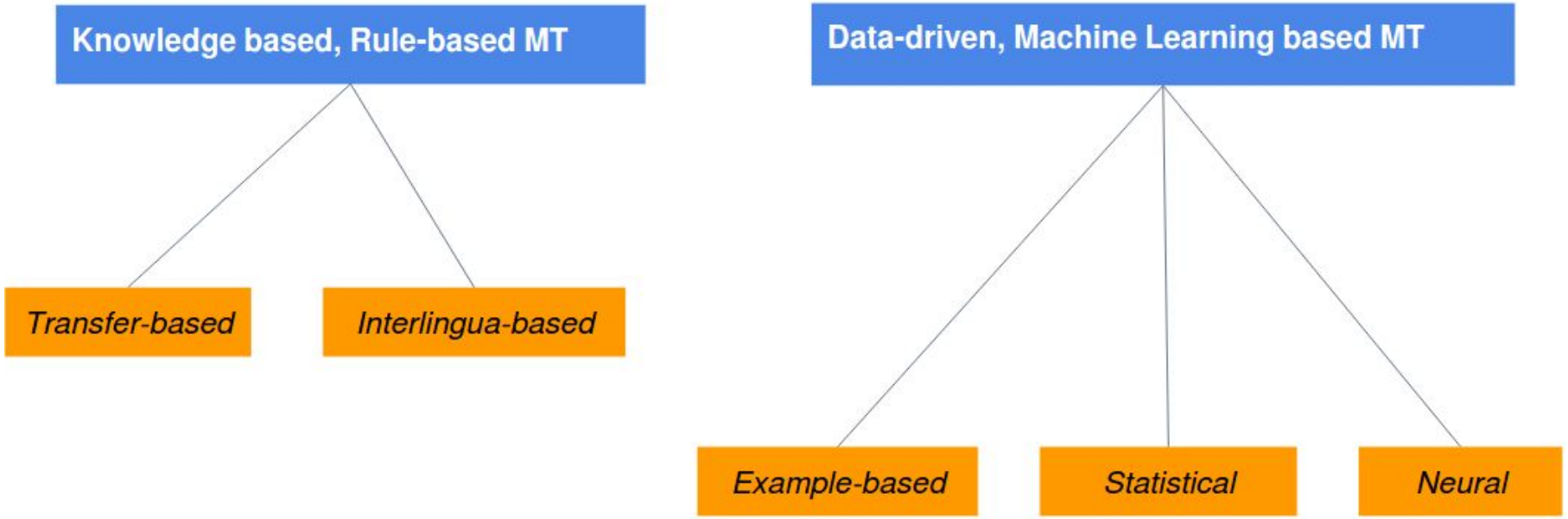
Available MT systems (1/2)

- Google Translate (https://translate.google.com/) supports Hindi, Bengali, Marathi, Punjabi, Gujarati, Tamil, Telugu, Kannada, Malayalam, Oriya,
- Microsoft's Bing Translator (https://www.bing.com/translator) supports Hindi, Bengali, Marathi, Punjabi, Gujarati, Tamil, Telugu, Kannada, Malayalam
- Facebook Translate (http://www.facebook.com/translations) supports Hindi, Bengali, Marathi, Punjabi, Gujarati, Tamil, Telugu, Kannada, Malayalam, Oriya
- Yandex Translate (https://translate.yandex.com/) supports Hindi, Bengali, Marathi, Punjabi, Gujarati, Tamil, Telugu, Kannada, Malayalam

Available MT systems (2/2)

- Does not support Indian languages
  - Systran (https://www.systransoft.com/)
  - Lingua Custodia (https://www.linguacustodia.finance/en/home/)
  - BabelFish (https://www.babelfish.com/)
  - SDL (https://www.sdl.com/)
  - Kantan MT (https://www.kantanmt.com/index.php)

Machine Translation: Approaches



MT Paradigms: 4 **A**s

- **Rule-based MT**
  - **A**nalysis
- **EBMT**
  - **A**nalogy
- **SMT**
  - **A**lignment
- **NMT**
  - **A**ttention

MT Paradigms

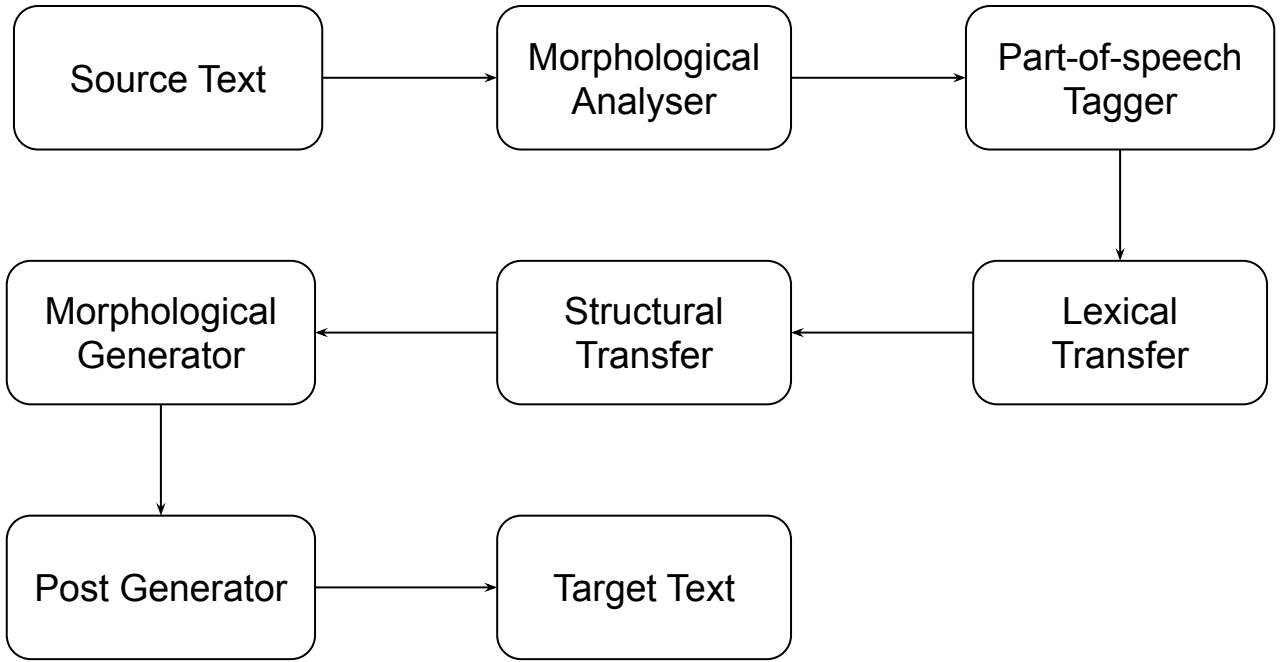
- **Rule-based MT**
  - Manually program lexicons/rules
  - SYSTRAN (AltaVista Babelfish; Originally from 70s)
- **Example-based MT**
  - Pattern based matching
- **Statistical MT**
  - Word-to-word, phrase-to-phrase probs
  - Learns translation rules from data, search for high-scoring translation outputs
    - Phrase or syntactic transformations
  - Key research in the early 90s: Google Translate (mid 00s)
  - Open-source: Moses
- **NMT**
  - Very recently deployed
  - Latent representations of words/phrases

Rule-based Machine Translation (RBMT)

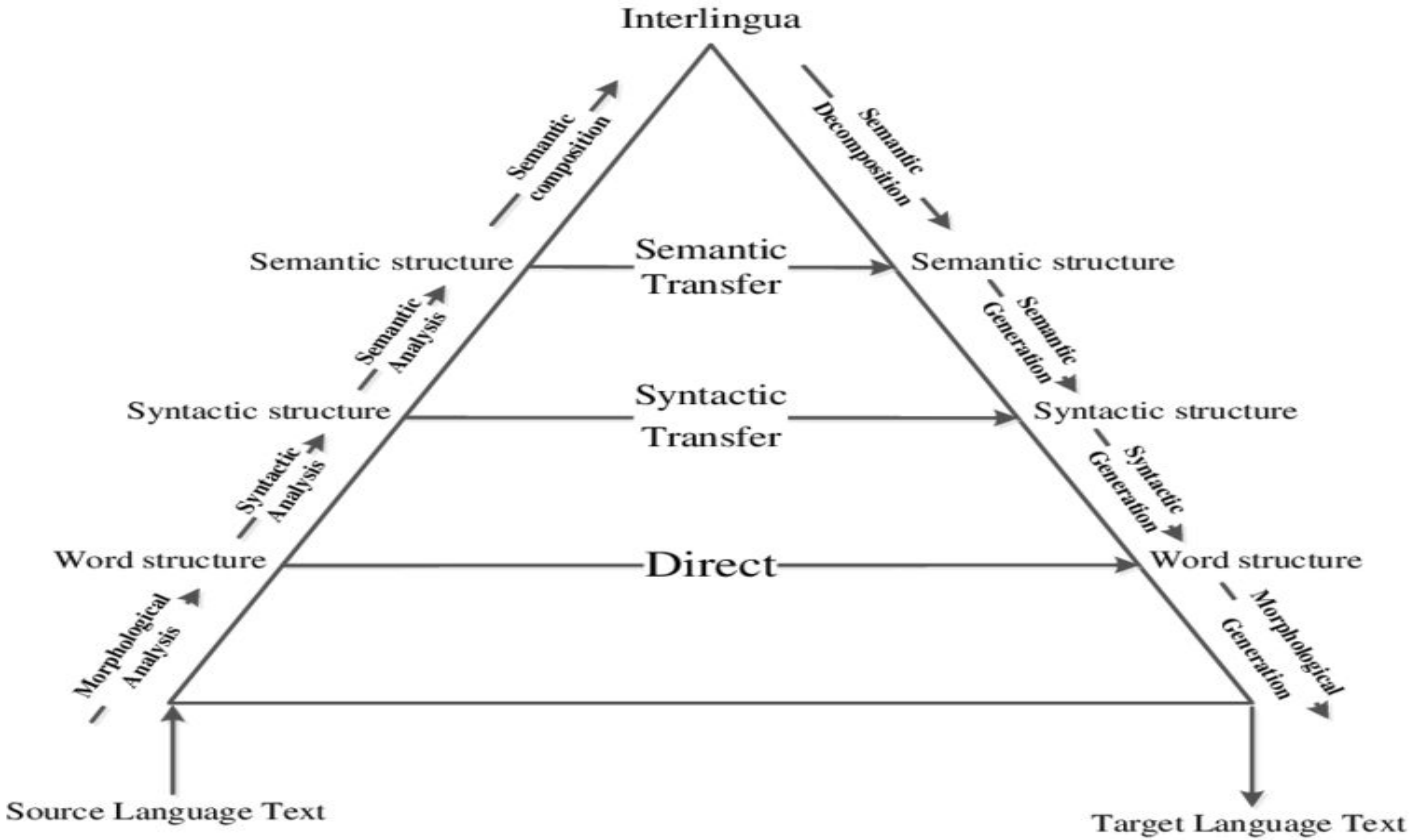
- Translation can be seen as three step process
  - **Analysis (A)**- Analyse the source language
  - **Transfer (T)**-Transfer source features into target language
  - **Generation (G)**-Generate the target sentence

- In RBMT, human created rules are used at every state of the A-T-G process

RBMT System Pipeline



The Vauquois Triangle



Example based Machine Translation

Example based Machine Translation-EBMT

- Corpus based translation approach
- Parallel sentences required (example database)
- Sentences are fragmented based on bilingual dictionary or by textual alignment
- Match input fragments into the example database (text similarity)
- Alignment information is extracted from text similarity measures (e.g: ***co-occurrence matrix***)
- Find corresponding aligned translated fragment using alignment information
- Recombine target fragments into output text

EBMT-An Example

- **Input:** He buys a book on international politics

Corpus matches:

- He buys a book — I read a book on international politics

वह एक किताब खरीदता है — मैंने अंतरराष्ट्रीय राजनीति पर एक किताब पढ़ी

- **Translation:** वह अंतरराष्ट्रीय राजनीति पर एक किताब खरीदता है

Statistical Machine Translation (SMT)

Statistical Machine Translation (SMT)

Three main components:

- **Phrase Table**: produces translation options and their probabilities for "phrases" (sequences of words) on the source language
- **Reordering Table**: indicates how words can be reordered when transferred from source language to target language
- **Language Model:** A language model which gives probability for each possible word sequence in the target language

IBM Models

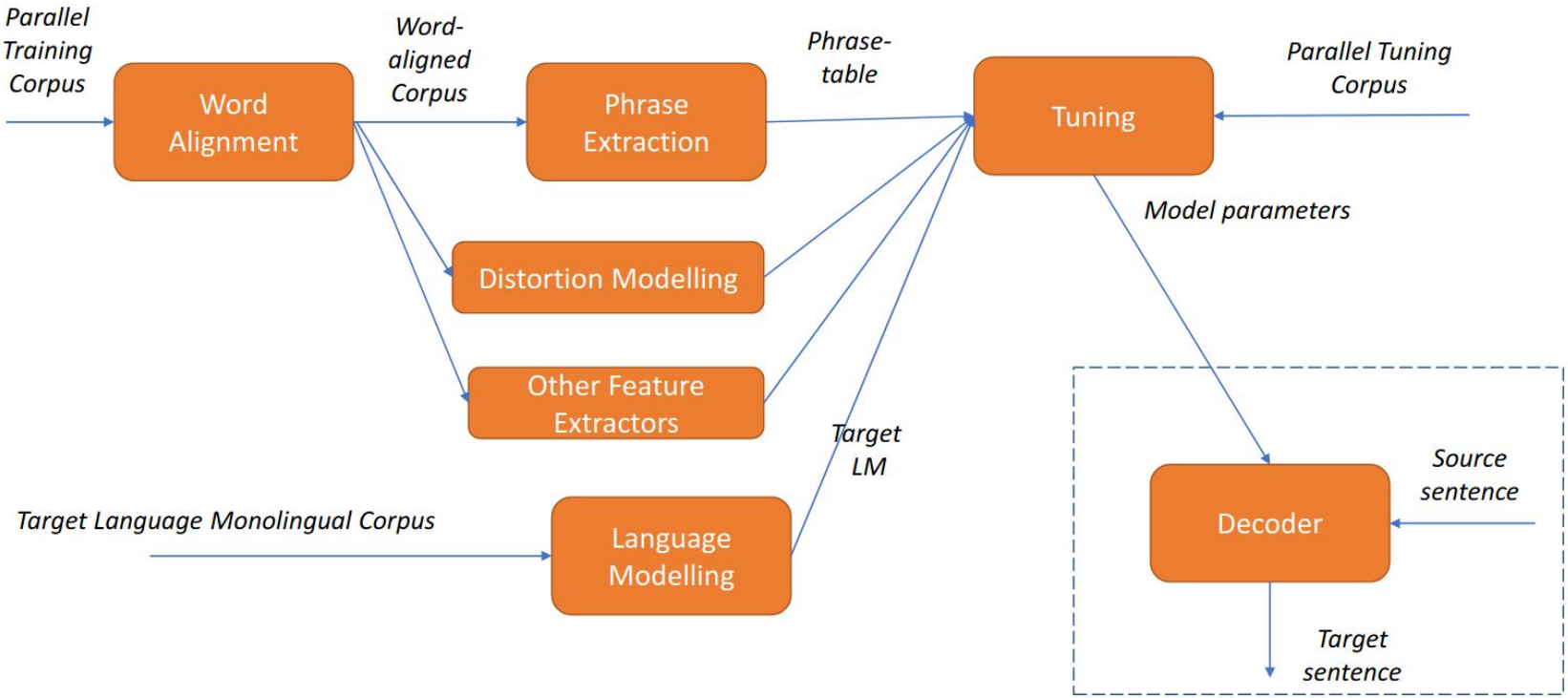
- IBM Model 1: Lexical translation
- IBM Model 2: Additional absolute alignment model
- IBM Model 3: Extra fertility model
- IBM Model 4: Added relative alignment model
- IBM Model 5: Fixed deficiency problem

IBM Model 1

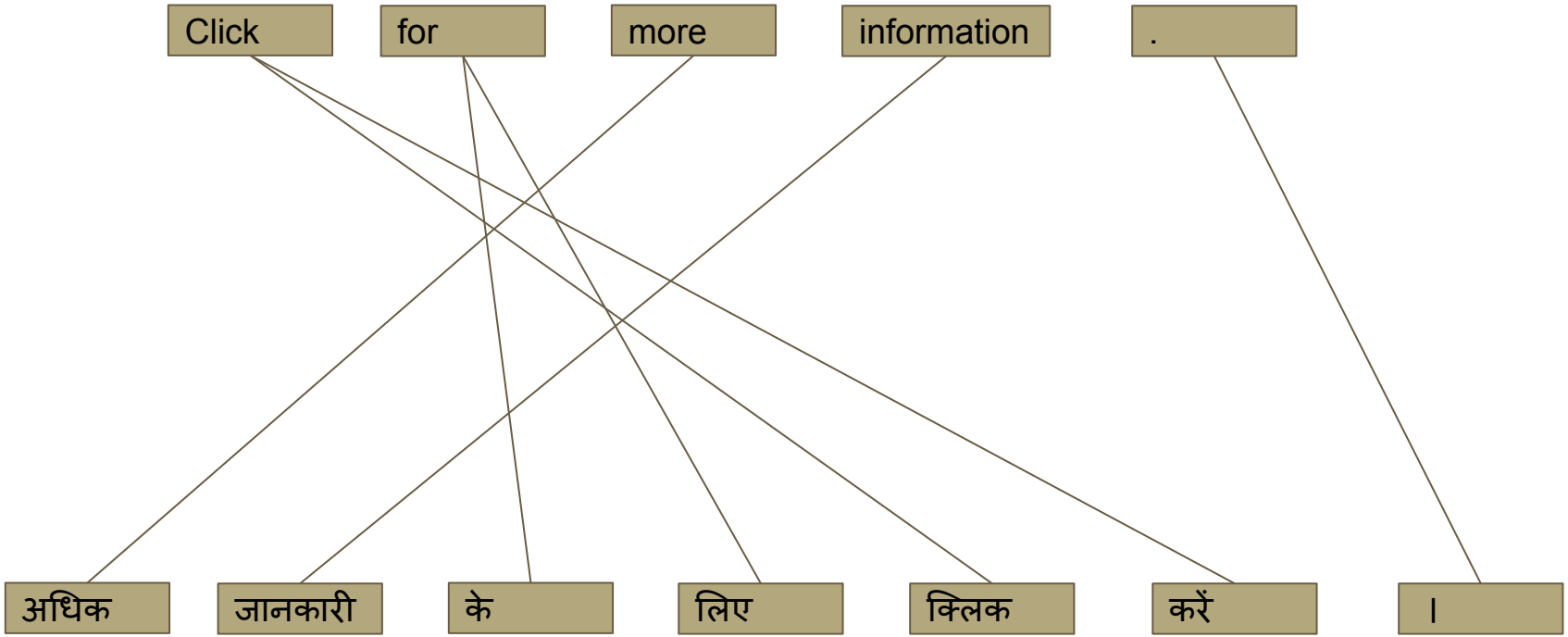
- **Generative model:** break up the translation process into smaller steps
  - IBM Model 1 only uses lexical translation
- **Translation probability**
  - Foreign sentence **f = (f_1, f_2, …, f_{lf})** of length **l_f**
  - Target English sentence **e = (e_1, e_2, …, e_{le})** of length **l_e**
  - With an alignment of each English word **e_j** to a foreign word **f_i** according to the alignment function **a: j → i**
  - Parameter ε is a normalization constant

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

SMT Pipeline



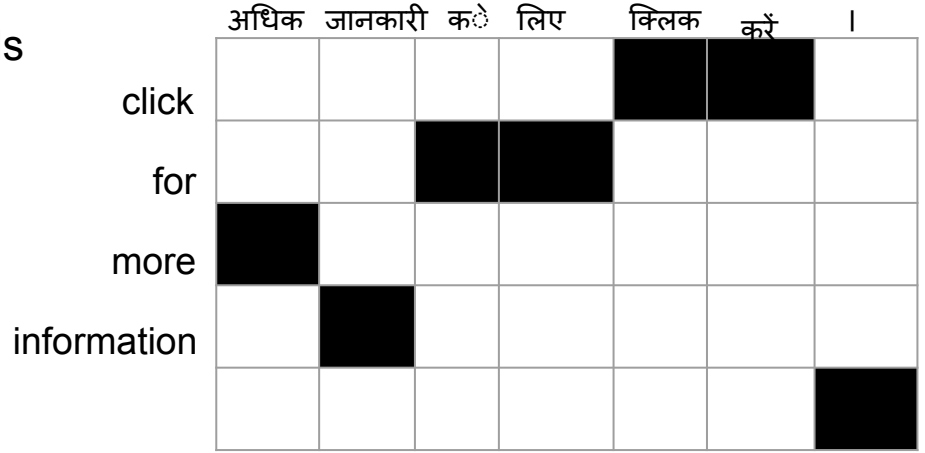
Word Alignment



Learning Phrase Tables using Word Alignments

- Starts with word alignment
- Consecutive sequence of aligned words makes a “phrase pair”

.



Learning Phrase Tables using Word Alignments

more - अधिक

Information - जानकारी

click - क्लिक करें

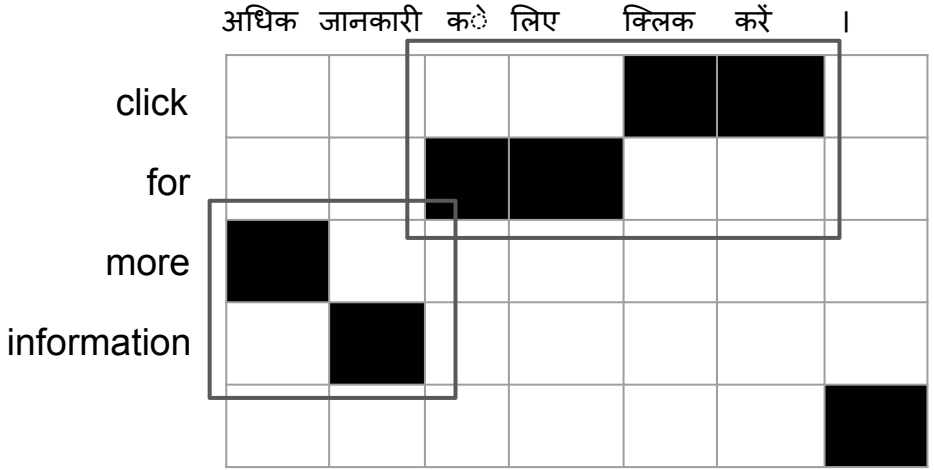
for - के लिए

. - ।

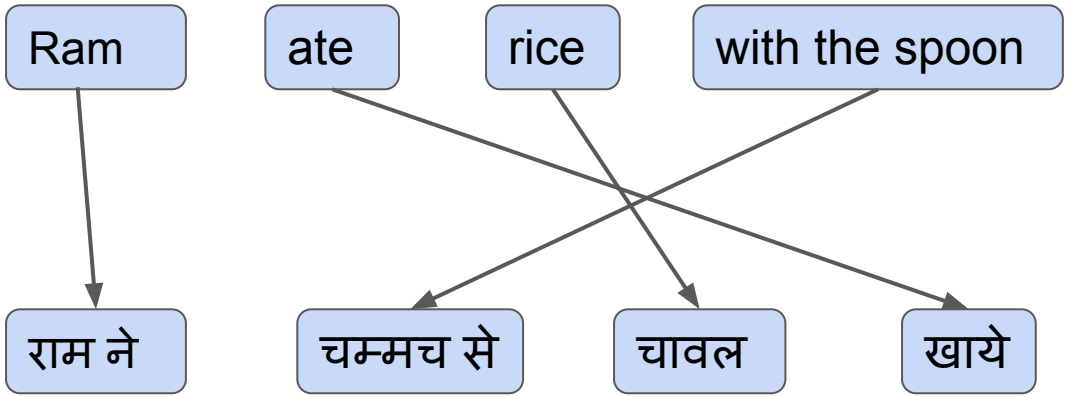
more information - अधिक जानकारी

click for - के लिए क्लिक करें

for more information - अधिक जानकारी के लिए



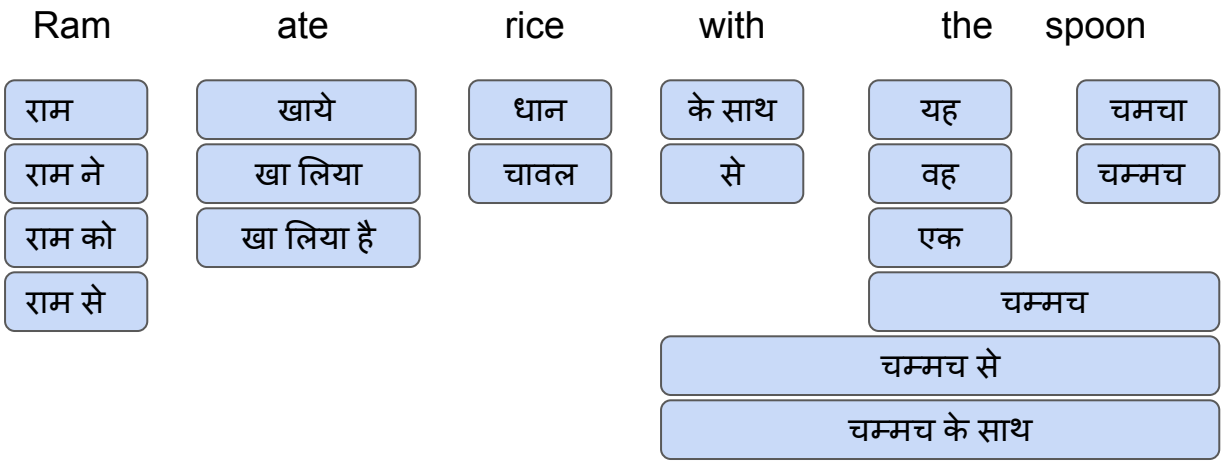
Decoding



- *Searching for the best translations in the space of all translations*

Decoding

- Sensible phrase translation is picked
- Phrase table may give multiple options to translate the input sentence
- Multiple possible word orders



Neural Machine Translation (NMT)

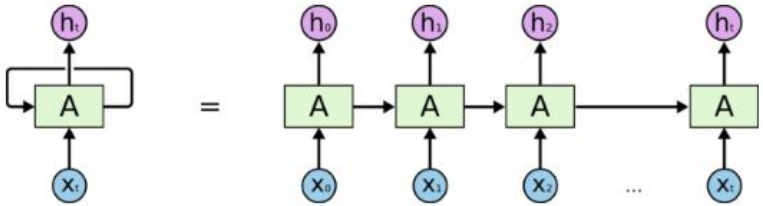
Neural Machine Translation (NMT)

- Neural Machine Translation: the MOST used technique today
- Deep neural network based

- Unlike PB-SMT, NMT does not translate piecewise

Encoder-Decoder Architecture

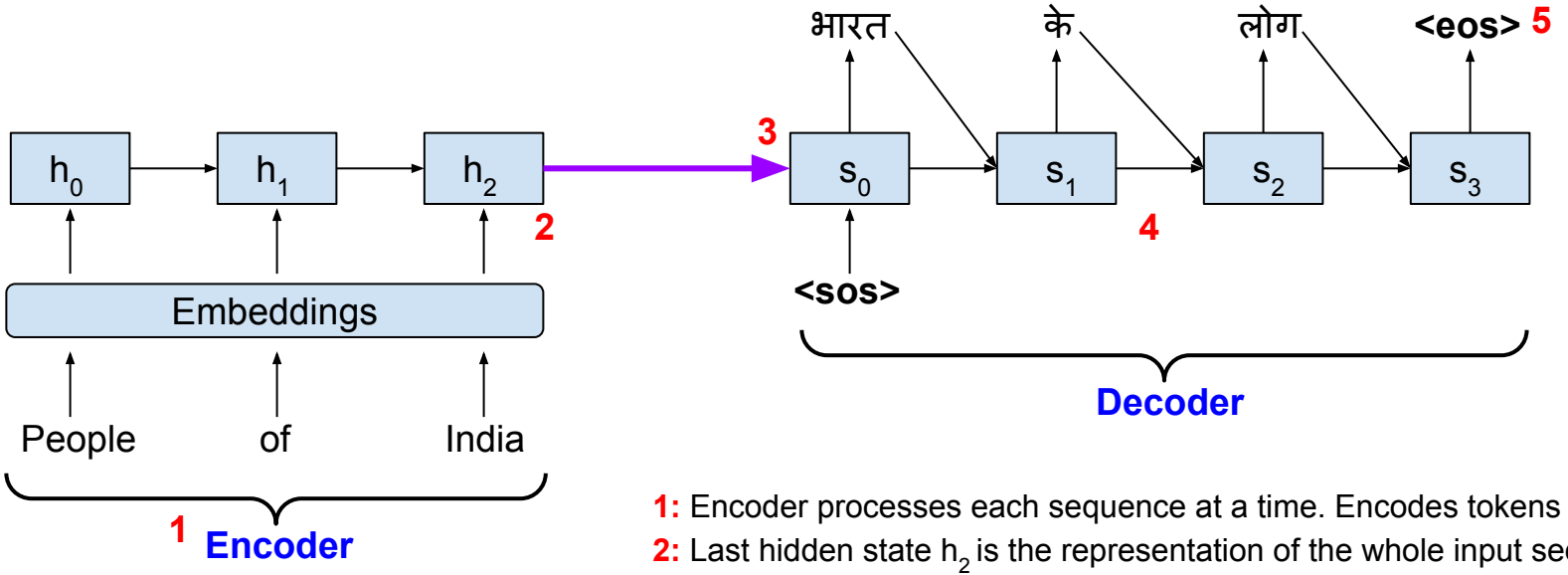
- Unlike PBSMT, whole sequence (sentence) is processed and generated in neural machine translation (NMT)
- Input and output would be the sequence of tokens (words) processed one at a time
  - Network (encoder and decoder) keeps information of what it has seen so far in the sequence while processing a token (word)
  - Recurrent neural network (RNN) is used to build encoder and decoder network
  - RNN uses the previous information while processing the current token



An unrolled recurrent neural network.

- https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Encoder-Decoder Architecture (NMT)



1: Encoder processes each sequence at a time. Encodes tokens one by one.

2: Last hidden state h_2 is the representation of the whole input sequence

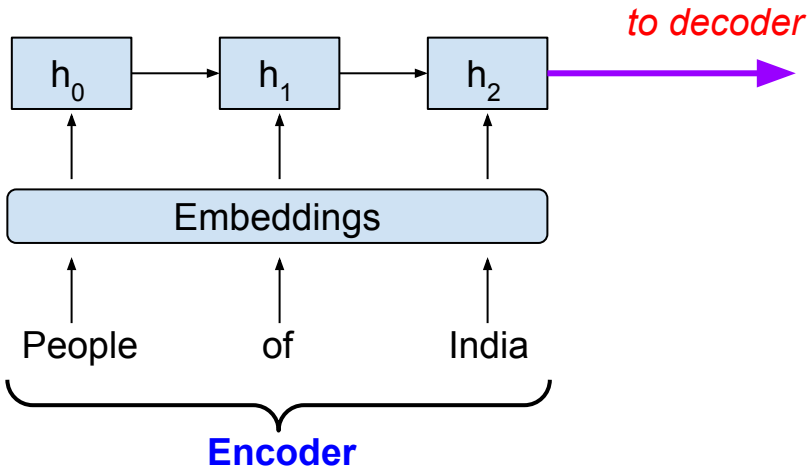
3: Initializing the decoder state. **<sos>** is special symbol to *start of sequence*

4: Decoder generates one token at a time from left to right

5: Decoding stops when **<eos>** is generated. **<eos>** represents *end of sequence*

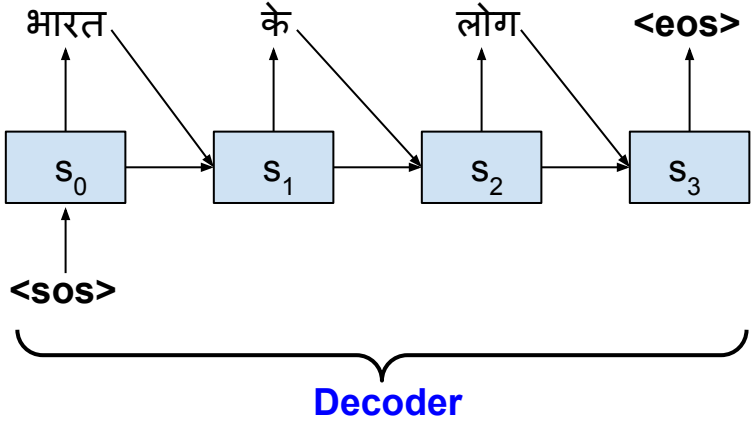
Encoder

- Encoder's hidden state:
  - $h_i = f(h_{i-1}, x_i)$
  - h_i is current hidden state
  - h_{i-1} is previous hidden state
  - x_i is current input



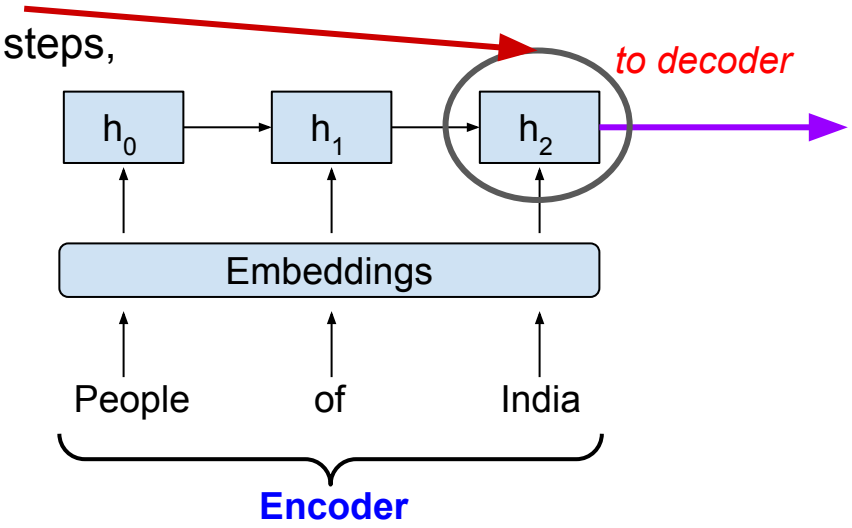
Decoder

- Output tokens:
  - $y_0, y_1, y_2, \ldots y_t$
- Decoder hidden states:
  - $s_0, s_1, s_2, \ldots s_t$
  - $s_i = g(s_{i-1}, y_{i-1}, h_x)$
  - h_x is encoder's last hidden state



Problem with Encoder-Decoder paradigm

- Encodes the entire sentence into a single vector
- Decoder uses this vector to generate output
- Due to long term dependency, after few time steps, sentence representation may get distorted



Solution

- Present the *relevant source sentence representation* for prediction at each step

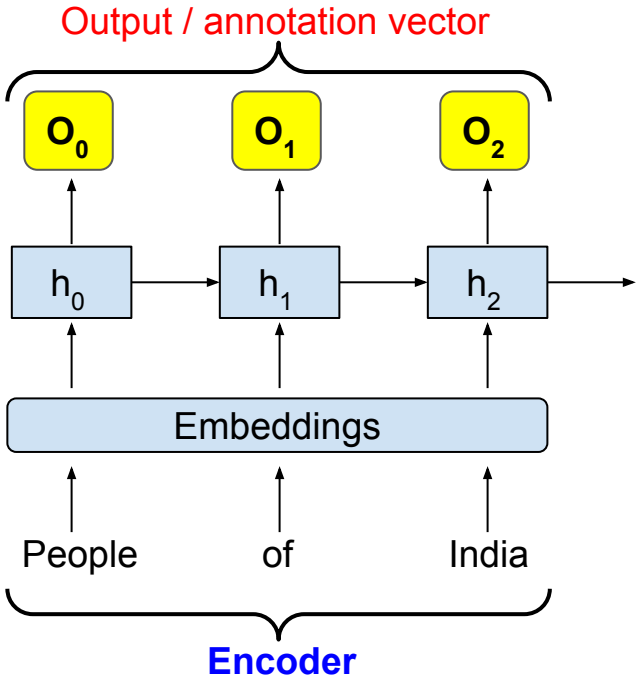
- *Apply attention to identify the significant source hidden states*

Encoder -- Attention -- Decoder paradigm

Annotation vectors

- Represent the source sentence by the set of output vectors from the encoder
- Encoder's output vectors are annotation vectors
  - Note: in the encoder-decode paradigm, we ignore the encoder outputs

- **Each output vector 'O_t' at time 't' is a contextual representation of the input at time 't'**



Use of annotation vectors by decoder

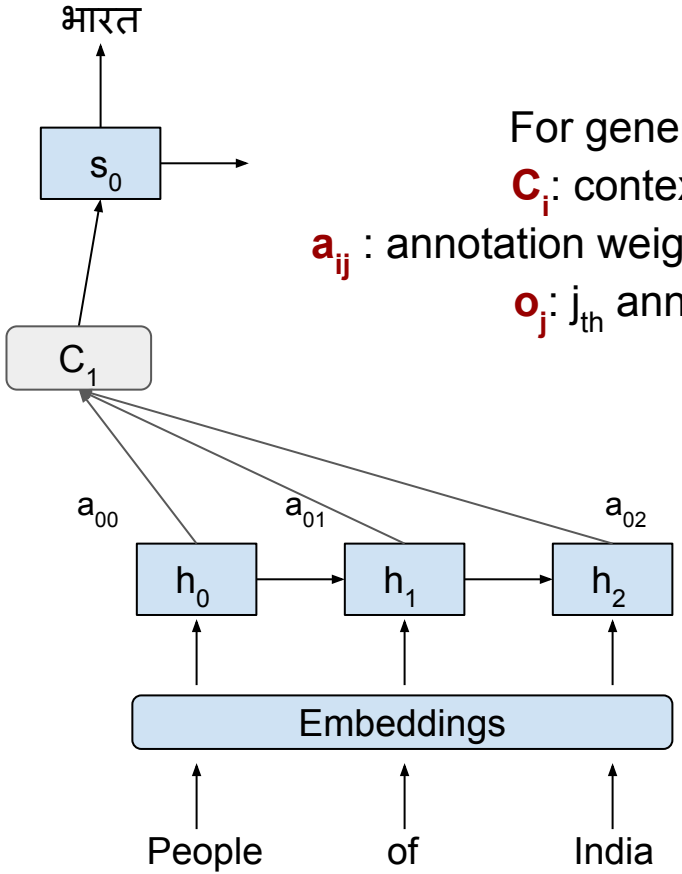
- **Not all annotation vectors are equally important for prediction of the next element**

- The annotation vector to use next depends on what has been generated so far by the decoder
- **One way to achieve this**: Take a weighted average of the annotation vectors, with more weight to annotation vectors which need more focus or attention

- ***This averaged context vector is an input to the decoder***

Attention

Attention based Enc-Dec



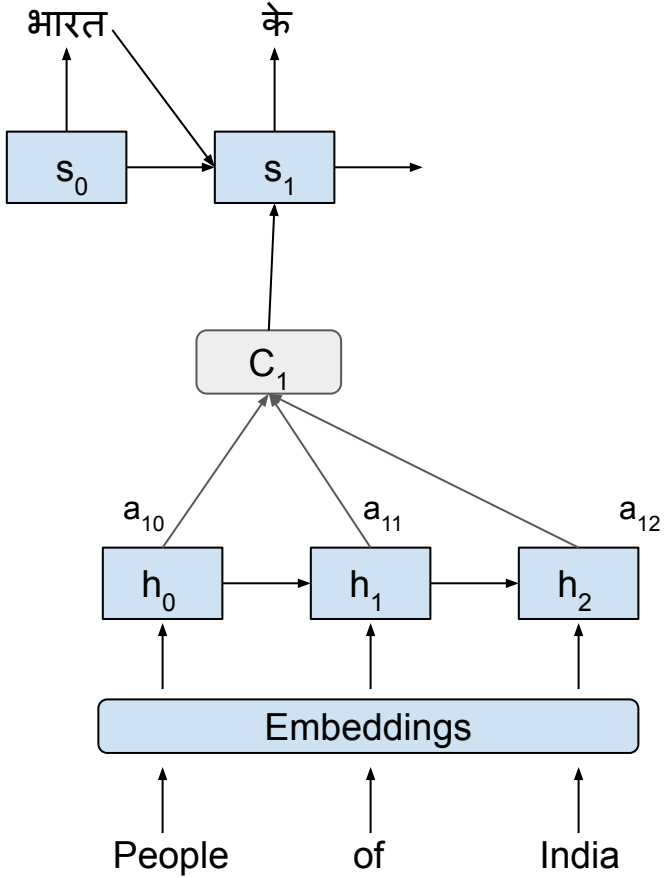
For generation of i_{th} output word:

$\mathbf{C_i}$: context vector

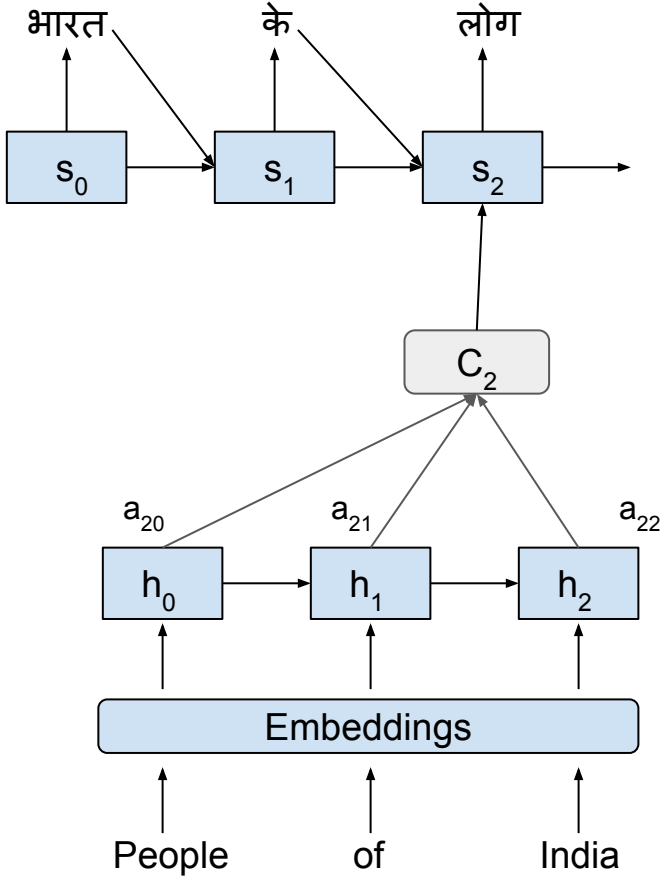
$\mathbf{a_{ij}}$: annotation weight for the j_{th} annotation vector

$\mathbf{o_j}$: j_{th} annotation vector

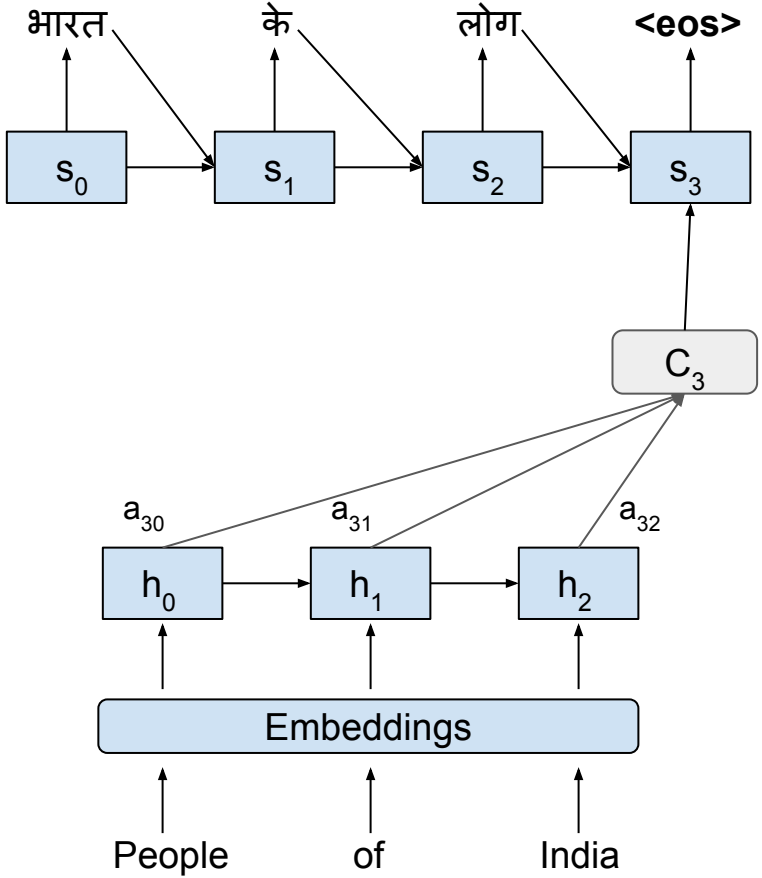
Attention



Attention


भारत
के
लोग
s_0
s_1
s_2
C_2
a_{20}
a_{21}
a_{22}
h_0
h_1
h_2
Embeddings
People
of
India

Attention


भारत
के
लोग
<eos>
s_0
s_1
s_2
s_3
C_3
a_{30}
a_{31}
a_{32}
h_0
h_1
h_2
Embeddings
People
of
India

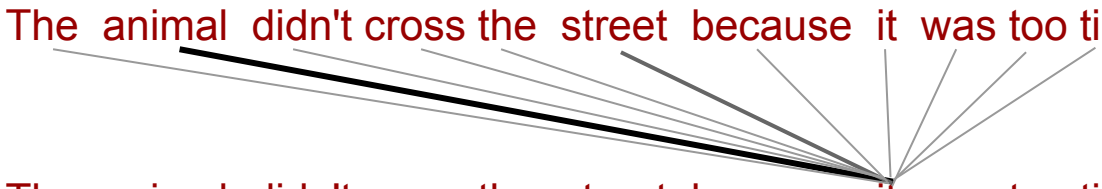
Transformer

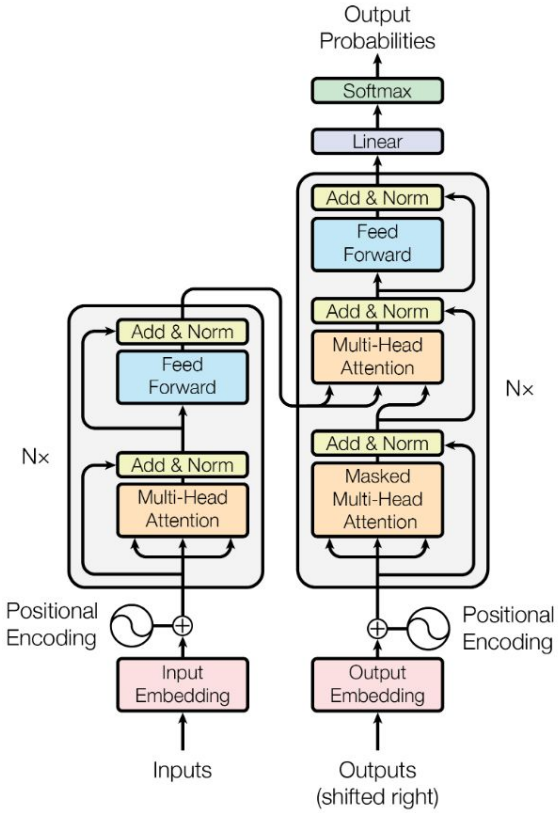
- “Attention is all you need”, Vaswani et al., NIPS 2017
- RNN generates hidden state as a function of previous state and current input sequentially
- Sequential nature prevents parallelization within training samples, especially when long dependency is available
- CNN reduces sequential nature but number of operations increases as the length between the input and output token increases
- Objective is to improve the performance and get rid of sequential computation

The Transformer uses multiple attention heads

Transformer

- **Self-attention layer:** a layer that helps the encoder look at other words within the input sentence as it encodes a specific word
- At the time of encoding a ***word***, self-attention helps to see that what other words within the sentence are related and significant to that **word**
- **Examples: “**It” word is more related to “animal” than “street” and other tokens.



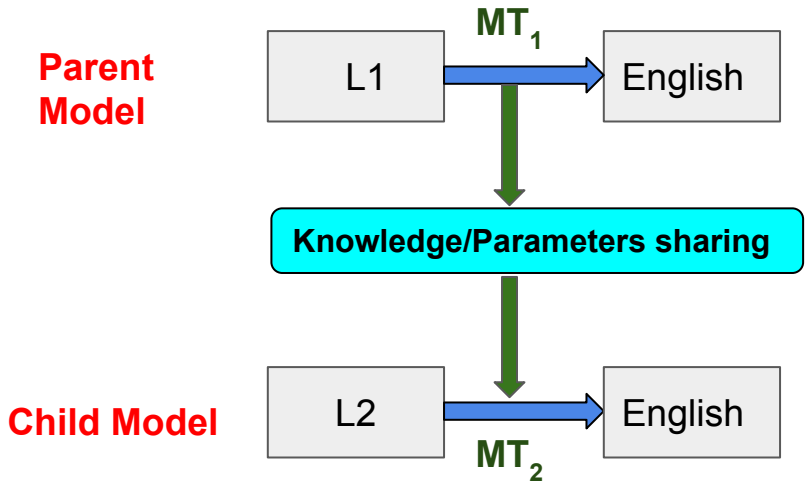


MT in Low-resources languages

- **Will cover some popular approaches**
  - Transfer Learning based MT
  - Pivot-based MT
  - Back-translation
  - Zero-shot MT
  - Domain Adaptation
  - Subword unit based MT
  - Unsupervised MT
  - Phrase based & Unsupervised Neural MT

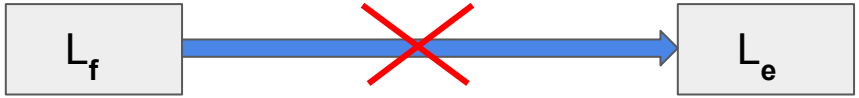
Transfer Learning

- Learning of a new task relies on the previously learned related task(s)
  - Learning is faster and requires less data
- **Knowledge transfer** from **high-resource** language pair to the **low-resource** language pair

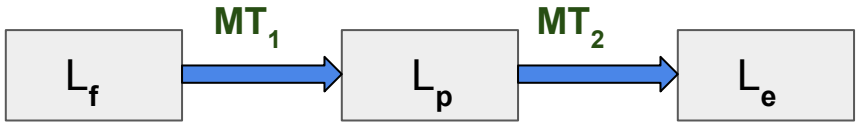


Pivot-based Machine Translation

- Used in absence of parallel corpus for a language pair ***(Source-Target)***
- A pivot language is used for translation
  - *Source → Pivot*
  - *Pivot → Target*
- ***Source-Pivot and Pivot-Target models are distantly related***



No parallel corpus. No direct MT System



L1 -to- L2 translation through 'L_p' PIVOT language

Transfer Learning for Low-Resource Neural Machine Translation

- Very first study to apply transfer learning to NMT
- English word embeddings from **parent model** (***French ⇒ English***) are copied to **child model** (***Uzbek ⇒ English***)

- Uzbek words are initially mapped to random French embeddings
- In child model, the *parameters of the English embeddings are frozen*, while the Uzbek *embeddings' parameters are allowed to be modified,* i.e. fine-tuned

***Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, 1568–1575. https://doi.org/10.18653/v1/D16-1163**

Transfer Learning

Language Pair: Uzbek-English (child model) and French-English (parent model)

Training Data Size:

	English side tokens
French-English	300 m Tokens
Uzbek-English	1.8m Tokens

BLEU Score:

Uzbek-English	Baseline	10.7
Uzbek-English	Transfer Learning	**15.0**

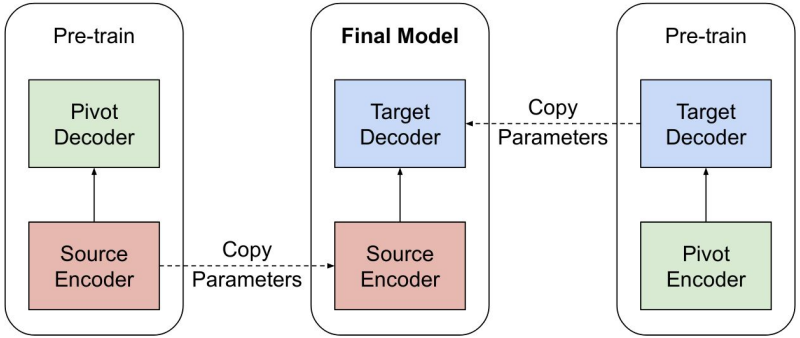
Pivot + Transfer Learning for NMT between Non-English Languages (1/6)

- Pre-train the **source-pivot** and **pivot-target** models

- To share the parameters and learn the final **source-target** model, 3 pivot based approaches were proposed

*Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 866–876. https://doi.org/10.18653/v1/D19-1080

Pivot + Transfer Learning: Plain Transfer Learning



- **Classical Transfer Learning:**
  - Pre-train a **source→pivot** model with a source-pivot parallel corpus
  - Pre-train a **pivot→target** model with a pivot-target parallel corpus
  - Initialize the source→target model with the source encoder from the pre-trained source→pivot model and the target decoder from the pre-trained pivot→target model
  - Continue the training with a **source-target** parallel corpus (***final model***)

- **Equivalent to zero-shot MT**
  - If the last step, i.e. training with source-target is skipped

*Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 866–876. https://doi.org/10.18653/v1/D19-1080

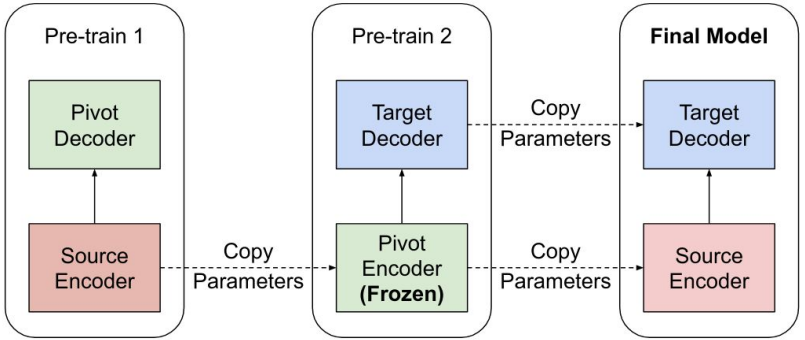
What are the problems with this pivot transfer learning?

- Source encoder is trained to be used by an English (i.e. pivot) decoder

- Target decoder is trained to use the outputs of an English (i.e. pivot) encoder and not of a **source encoder**

Hence, introduces the inconsistency of source-pivot and pivot-target pre-training stages

Pivot + Transfer Learning: Step-wise Pre-training



- **Step-wise pre-training**
  - Pivot encoder is initialized by the source enc.
  - Joint vocab of source and pivot sentences is built so that the pivot encoder effectively represents both languages
  - Freeze pre-trained (pivot) encoder for the source language in the first step, but also able to encode a pivot language sentence in a similar representation space
  - It is effective for linguistically similar languages
- **Freezing the pre-trained encoder ensures that**
  - Even after the second step, the encoder still models the source language although we train the NMT model for pivot-target
  - Without the freezing, the encoder completely adapts to the pivot language input

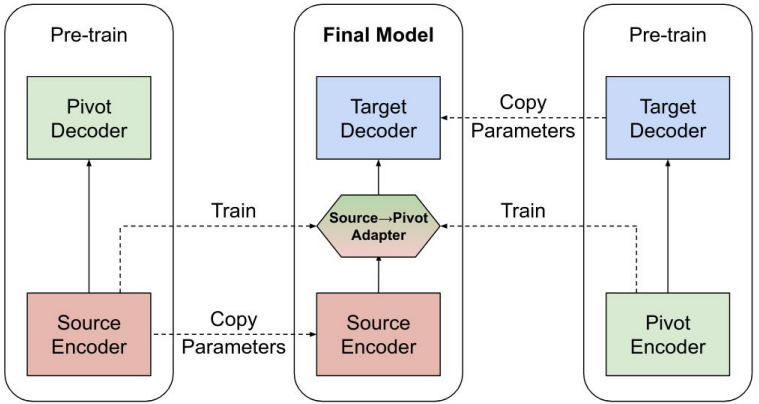
***Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Confe**rence on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 866–876. https://doi.org/10.18653/v1/D19-1080

Pivot + Transfer Learning: Pivot Adapter (1/2)

- Stage-wise pre-training can be avoided
- Postprocess the network to enhance the connection between the source encoder and the target decoder which are pre-trained individually

- After the pre-training steps, we adapt the **source encoder outputs** to the **pivot encoder outputs** to which the target decoder is more familiar
- Learn a linear mapping between the two representation spaces with a small source-pivot parallel corpus

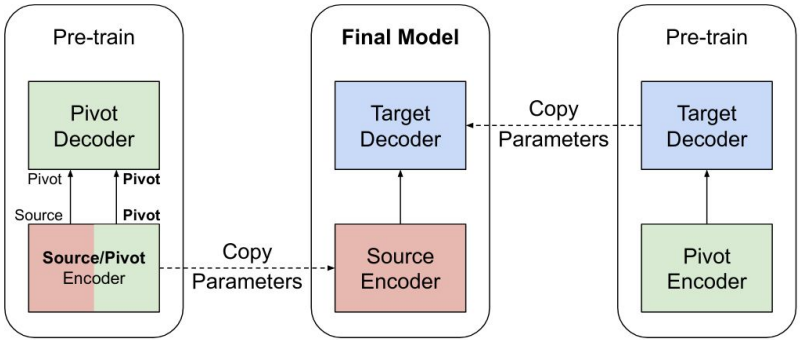
Pivot + Transfer Learning: Pivot Adapter (2/2)



- **Pivot Adapter**
  - Decoder of the final model is initialized by the decoder of **pivot-target** model
  - Encoder of the final model is initialized by the encoder of **source-pivot** model
  - A mapping M is learned to minimize the pooled representation of **s** (vector representation of source sentence) and **p** (vector representation of pivot sentence)
  - The learned mapping will be multiplied by the output of source⇒target output

Pivot + Transfer Learning:Cross-lingual Encoder

- **Cross-lingual Encoder**
  - Decoder of final model is initialized by the decoder of pivot-target model
  - Source-pivot model is trained for source-pivot and pivot-pivot (autoencoder) languages both



- Pivot decoder decodes pivot sentences for source input and pivot input
  - In source-pivot model noisy *(random token drop, token order shuffle etc.)* pivot input is used so that decoder can learn linguistic structure of pivot sentence too
  - Now the source encoder of source-target (final) model is initialized with source encoder of source-pivot model

*Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 866–876. https://doi.org/10.18653/v1/D19-1080

In a nutshell!

- Cross-lingual encoder changes the encoder pretraining stage (**source-pivot**)
- Step-wise pre-training modifies decoder pre-training stage (**pivot-target**)
- Pivot adapter is applied after all the pre-training steps

Pivot + Transfer Learning: Results

Language Pair:

a) Fr → En ===> En → De,

b) De → En ===> En → Cz

Usage	Data	Sentences
Pre-train	Fr-En En-De	35M 9.1M
Fine-tune	Fr-De	270k
Pre-train	De-En En-Cs	9.1M 49M
Fine-tune	De-Cs	230k

Training Data

	French ⇒ German (newstest13)	German ⇒ Czech (newstest13)
Direct translation	16.0	12.8
Plain transfer	18.7	18.0
+pivot adapter	19.1	18.7
+ cross-lingual encoder	18.9	17.6
+ pivot adapter	19.1	18.1
Step-wise pre-training	19.9	18.1
+ cross-lingual encoder	20.7	19.1

Bleu Score

Back-translation

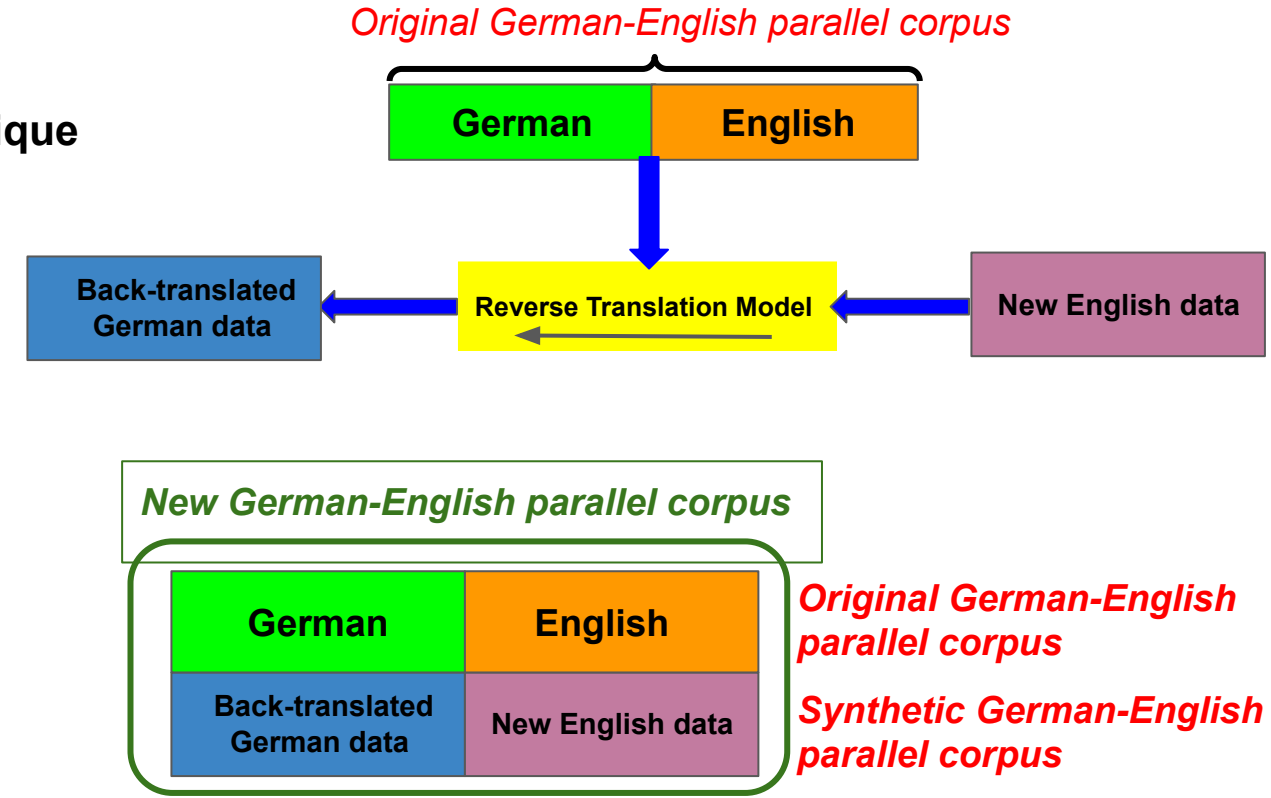
- Objective is to improve NMT model by incorporating ***target side monolingual data***
- Instead of using ***RNN language model*** *trained on* ***target monolingual data*** at decoder side, incorporate monolingual sentences with available training data only

- **Synthetic parallel corpus**
  - With the available parallel corpus, **target ⇒ source** model is trained and used to translate target monolingual data into source language
  - **Original + synthetic parallel corpus** is used to train final **source ⇒ target** model

***Sennrich, Rico, Barry Haddow and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data." *ACL 2016*.**

Back-translation

A data enrichment technique



Back-translation

- **Training data size:**

	Parallel Data	Synthetic data (parallel)
German→English	4,200,000	3,600,000 (En→De)
Turkish→English	320,000	3,200,000 (En→Tu)

- **Bleu Score:**

	German→ English	Turkish→ English
Baseline	26.7	18.8
+Synthetic	**30.4**	**21.2**

Domain Adaptation

- Needs in a domain-specific **low-resource** scenario
- Improves the translation quality for a specific domain

- Less data in specific domains (e.g. Judicial domain)
  - NMT model trained on large **out-domain** parallel data for a specific pair
  - Fine-tune using limited **in-domain** parallel data

Let us see

How can we train a translation model on multi-domain data to improve the test-time performance in each constituent domain?

Effective Domain Mixing for Neural Machine Translation

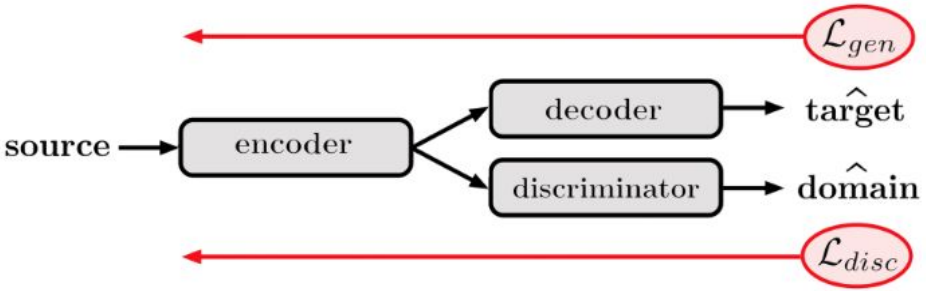
- Mixing data from heterogeneous domains degrades performance of the model compared to single-domain setting
- Proposed models do not require knowledge of each example's domain at inference time
- Adversarial learning mechanism forces the encoder to encode relevant information regarding specific domain

- **Contributions**
  - Discriminative mixing
  - Adversarial discriminative mixing
  - Target-side token mixing
- Dataset '$\mathcal{D}$' consists of source sequences 'X' , target sequences 'Y' and domain class labels 'D' that are only known at the training time

*Britz, D., Le, Q. and Pryzant, R., 2017, September. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation* (pp. 118-126).

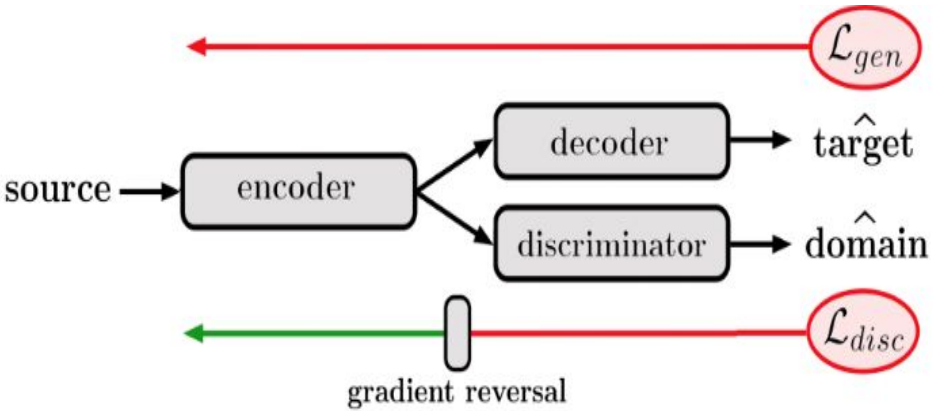
Effective Domain Mixing for Neural Machine Translation: Discriminative Mixing

- Discriminator network is added on top of encoder
- It takes single vector encoding of source as input and predicts the correct domain
- **This forces encoder to encode domain relevant information**
- Encoder and decoder are LSTM based networks
- Discriminator is a fully connected feed-forward network



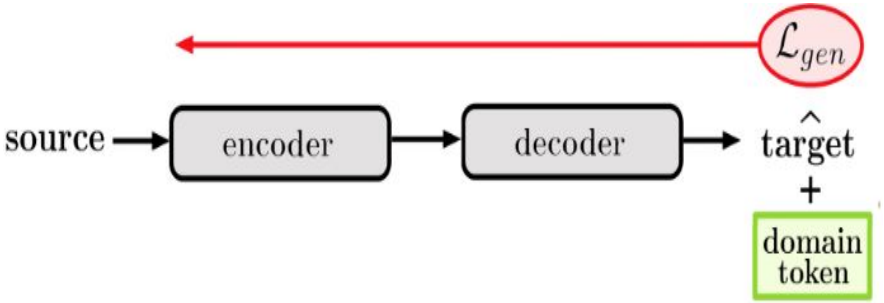
Effective Domain Mixing for Neural Machine Translation: **Adversarial Discriminative Mixing**

- Similar to that of discriminative mixing but gradients from discriminator network to encoder are reversed
- It produces opposite effect of discriminative mixing
- Discriminator still learns to distinguish between domains
- **Encoder forced to compute domain invariant representations that are not useful to the discriminator**



Effective domain mixing for Neural Machine Translation:**Target token mixing**

- Prepending a domain token to the target sequence
- For example, the domain token can be "**domain=judicial**"



- Decoder must learn to predict the correct domain token based on the source at first step of decoding
- Technique has similar regularizing effect as adding a discriminator network

Dataset Statistics

Language pair	Domain	Size of data
English - Japanese	ASPEC and Subcrawl	3m
English - Chinese	News and TED	200k
English - French	Europarl and OpenSubs	200k

Evaluation Results (1/2)

EN - JA Model	ASPEC	SubCrawl
ASPEC	38.87	3.85
SubCrawl	2.74	16.91
ASPEC + SubCrawl	33.85	14.34
Discriminator	35.01	**15.38**
Adv. Discriminator	29.87	13.31
Target Token	**35.05**	14.92

EN - ZH Model	News	TED
News	12.75	3.12
TED	2.79	8.41
News + TED	11.36	6.67
Discriminator	**12.88**	**8.64**
Adv. Discriminator	12.15	8.16
Target Token	11.98	7.69

Row-Training domain; **Column**- Test domain

Proposed model performs better for mixed domain

Evaluation Results (2/2)

EN - FR Model	Europarl	OpenSubs
Europarl	34.51	13.36
OpenSubs	13.12	15.20
Europarl + OpenSubs	38.26	27.90
Discriminator	39.03	**27.91**
Adv. Discriminator	38.38	25.67
Target Token	**39.10**	25.32

Subword NMT (1/7)

- **Subword**: An effective way to handle out-of-vocabulary (OOV) word problem in NMT
  - MT is open vocabulary problem but NMT works with a predefined vocabulary size
  - Typical vocab size: 50k words, where each of remaining words/tokens is treated as unknown (UNK)
- Motivated by byte-pair-encoding
- In subwordification, a word is splitted into multiple subparts and each subpart (***may be meaning-ful/less***) is considered as a separate token
- **E.g:**
  - Superstition = super@@ stition

    @@ is used to differentiate between a word and a subword in corpus

- All words may not be splitted into subwords

Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In ACL 2016

Subword NMT (2/7)

Step 1: Represent each word in the corpus as a combination of the characters along with the special end of word token </w>

For example, suppose the corpus has only four words (with counts): low:5, lower:2, newest:6, widest:3

Word	count
l o w </w>	5
l o w e r </w>	2
n e w e s t </w>	6
w i d e s t </w>	3

Vocabulary

l o w </w> e r n s t i d

Subword NMT (3/7)

Step 2: repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’

Stopping criteria: After predefined number of merge operations (i.e. step 2)

After 1 merge

Word	count
l o w </w>	5
l o w e r </w>	2
n e w **es** t </w>	6
w i d **es** t </w>	3

Vocabulary

l o w </w> e r n s t i d
es

Subword NMT (4/7)

Step 2: repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’

Stopping criteria: After predefined number of merge operations (i.e. step 2)

After 2 merges

Word	count
l o w </w>	5
l o w e r </w>	2
n e w **est** </w>	6
w i d **est** </w>	3

Vocabulary

l o w </w> e r n s t i d
es **est**

Subword NMT (5/7)

Step 2: repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’

Stopping criteria: After predefined number of merge operations (i.e. step 2)

After 3 merges

Word	count
l o w </w>	5
l o w e r </w>	2
n e w **est</w>**	6
w i d **est</w>**	3

Vocabulary

l o w </w> e r n s t i d
es est **est</w>**

Subword NMT (6/7)

Step 2: repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’

Stopping criteria: After predefined number of merge operations (i.e. step 2)

After 4 merges

Word	count
lo w </w>	5
lo w e r </w>	2
n e w est</w>	6
w i d est</w>	3

Vocabulary

l o w </w> e r n s t i d
es est est</w> **lo**

Subword NMT (7/7)

Step 2: repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’

Stopping criteria: After predefined number of merge operations (i.e. step 2)

Word	count
low </w>	5
low e r </w>	2
n e w est</w>	6
w i d est</w>	3

After 5 merges

Vocabulary

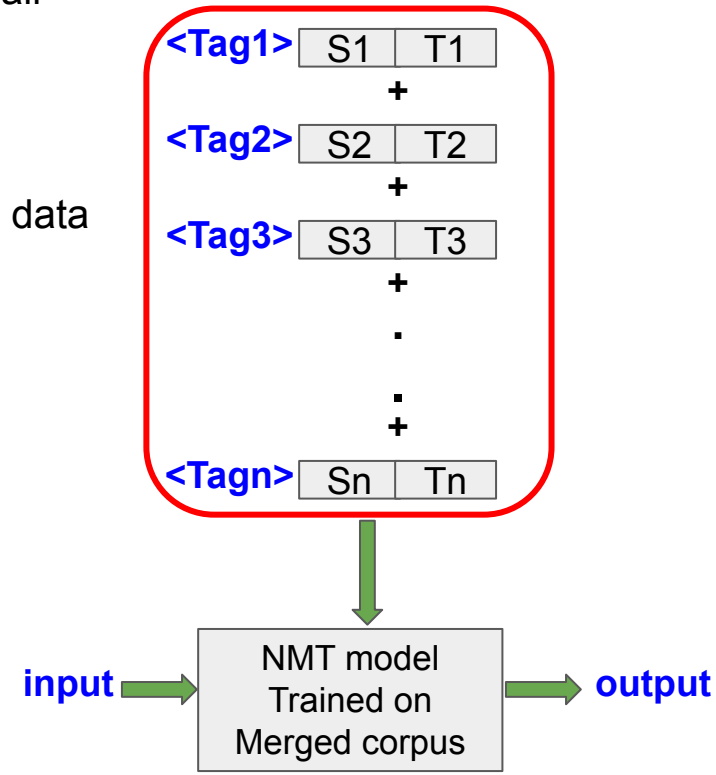
l o w </w> e r n s t i d
es est est</w> lo **low**

Subword NMT: Results

S.No	Model	Source vocabulary	Target vocabulary	BLEU
1	English to German - Word level	300000	500000	24.2
2	English to German - Subword	90000	90000	24.7
3	English to Russian - Word level	300000	500000	22.8
4	English to Russian - Subword	90000	100000	24.1

Google's Multilingual Neural Machine Translation System: *Enabling Zero-Shot Translation*

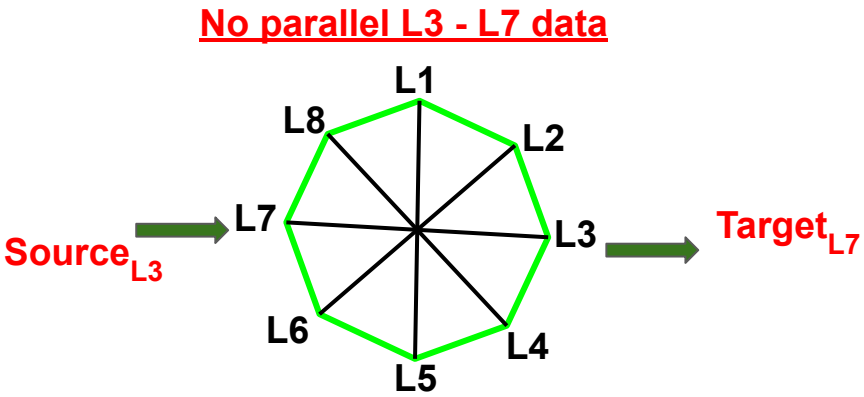
- An artificial tag is appended before each source-target pair
  - **<HI## People of India भारत के लोग>**
- All tagged sentences from different language pairs are merged
- Multilingual NMT model is trained on the tagged merged data
- **All the parameters are shared**
- **High-resource language pairs help low-resource language pairs**



*Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G. and Hughes, M., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, pp.339-351.

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

- **Zero Shot translation**
  - **Translation between language pairs for which the model is not trained directly**



Google's Multilingual Neural Machine Translation System: *Enabling Zero-Shot Translation*

WMT: WMT dataset, **Prod:** Google-internal large-scale production datasets

- **De:** German, **Fr:** French, **En:** English, **Ja:** Japanese, **Ko:** Korean **Pt:** Portuguese, **Es: Spanish**

Model	Single	Multi	Diff
WMT De→En	30.43	30.59	+0.16
WMT Fr→En	35.50	35.73	+0.23
WMT De→En⋆	30.43	30.54	+0.11
WMT Fr→En⋆	35.50	36.77	+1.27
Prod Ja→En	23.41	23.87	+0.46
Prod Ko→En	25.42	25.47	+0.05
Prod Es→En	38.00	38.73	+0.73
Prod Pt→En	44.40	45.19	+0.79

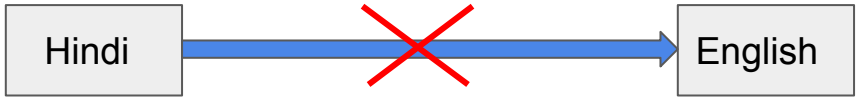
Many-to-One

Model	Single	Multi	Diff
WMT En→De	24.67	24.97	+0.30
WMT En→Fr	38.95	36.84	-2.11
WMT En→De⋆	24.67	22.61	-2.06
WMT En→Fr⋆	38.95	38.16	-0.79
Prod En→Ja	23.66	23.73	+0.07
Prod En→Ko	19.75	19.58	-0.17
Prod En→Es	34.50	35.40	+0.90
Prod En→Pt	38.40	38.63	+0.23

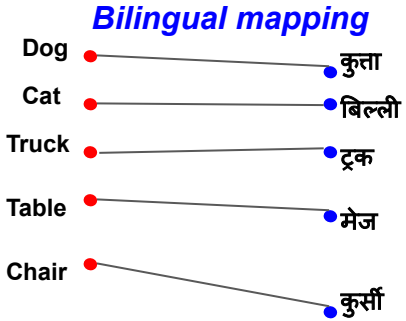
One-to-Many

Unsupervised Neural Machine Translation

- Supervised machine translation requires parallel sentences for training
- Unsupervised MT requires no parallel data
- Un-NMT ***requires aligned word embeddings***

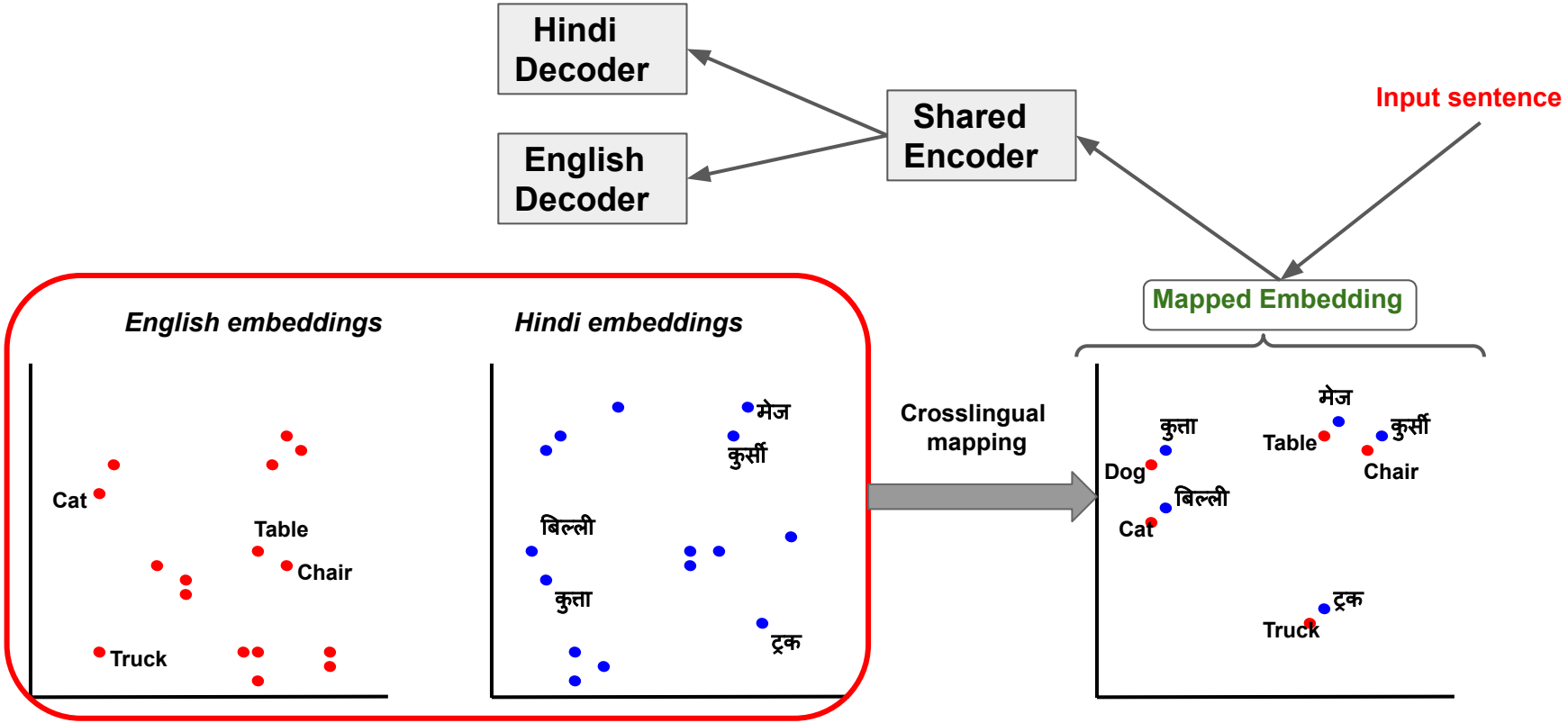


No parallel corpus. No Problem.



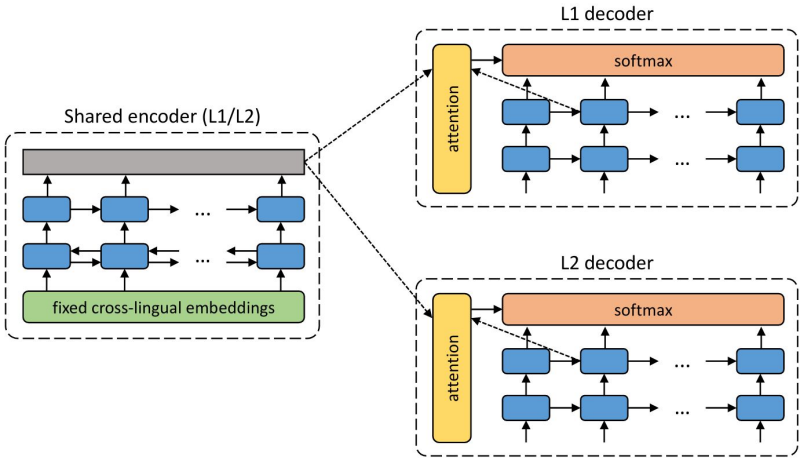
Artetxe, Mikel, Gorka Labaka, Eneko Agirre and Kyunghyun Cho (2018). Unsupervised Neural Machine Translation.ICLR

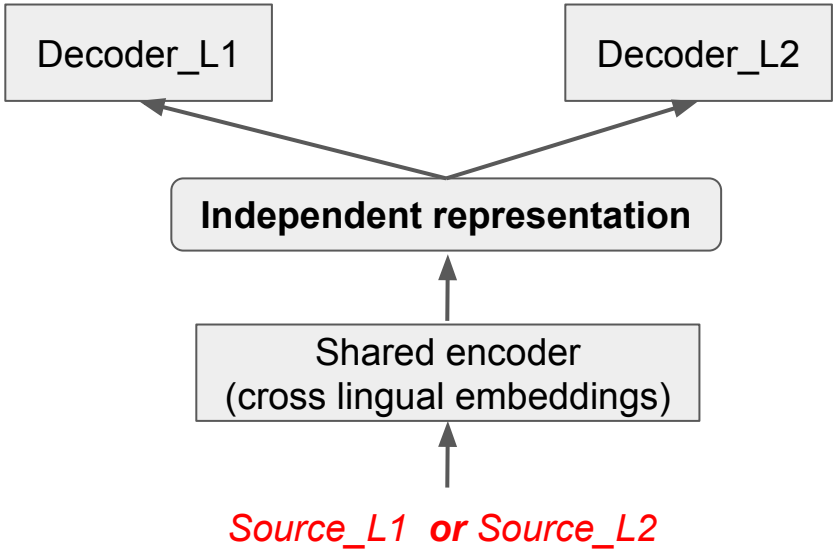
Cross-lingual embedding mapping



Unsupervised Neural Machine Translation

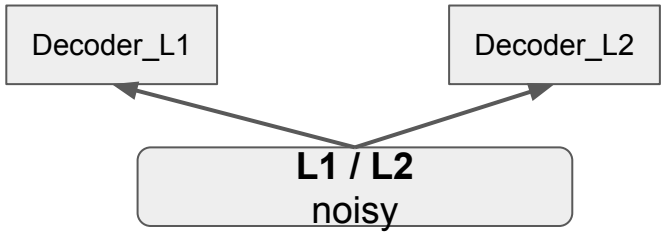
- Shared encoder is initialized by cross-lingual embedding which generates a language Independent representation which is used by language specific decoders



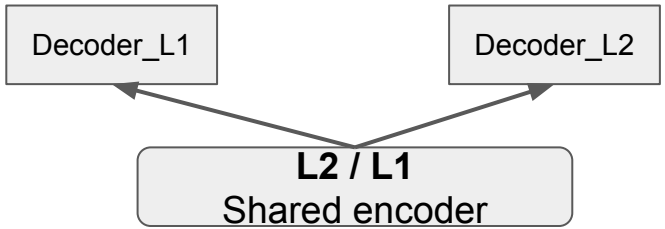


Unsupervised Neural Machine Translation

- Two main components of UNMT which executes in iterative manner (batch wise)
- **Denoising Autoencoder**
  - Same language's encoder-decoder used as auto-encoder. Takes noisy $input_{L1/L2}$ and $output_{L1/L2}$ decode using same language decoder
- **On-the-fly Back-translation**
  - Different language's encoder-decoder used to translate. Takes $input_{L1/L2}$ and decode as $output_{L2/L1}$ using different language's decoder



Denoising autoencoder



On-the-fly back-translation

Steps

- For each sentence in language L1, the system is trained alternating two steps: ***denoising*** and ***on-the-fly back translation***
- **Denoising:** optimizes the probability of encoding a noised version of L1 with the shared encoder and reconstructing it with the L1 decoder
- **On-the-fly back-translation**
  - Translates L1 into L2 in inference mode
    - Encodes it with the shared encoder (L1/L2)
    - Decodes with the L2 decoder
  - Optimizes the probability of encoding the translated sentence with the shared encoder
  - Recovers the original sentence with the L1 decoder

Training alternates between sentences in L1 and L2, with analogous steps for the latter

Data and Results

Monolingual training data size:

English	495m
French	122m
German	622m

BLEU Score:

	Fr→ En	**En→ Fr**	**De→ En**	**En→ De**
Baseline (word-by-word translation)	9.98	6.25	7.07	4.39
UNMT	**15.56**	**15.13**	**10.21**	**6.55**

Multilingual Unsupervised NMT: Background

- Unsupervised NMT (Artetxe et al., 2018)

- **Features**
  - Consists of a shared Encoder and two Decoders
  - Fixed cross-lingual embedding at encoder side
  - Denoising Auto-encoding
  - Back-translation

How is it different from Artetxe et al. (2018)?

- For **Multilingual NMT** by Artetxe et al (2018) (for n languages)
  - n denoisings
  - 2*n*(n-1) back-translations [*Each of n to n-1 and each of n-1 to n*]

- For **Multilingual NMT,** in our approach
  - n denoisings
  - 2*(n-1) back-translations [***L1 to all n-1; and each of n-1 to L1***]
- As a by-product, we generate ***zero-shot*** translation (*without any parallel corpus, even via any other language*)

Our approach

1. Joint training of multiple languages (L_1, L_2, L_3, ... , L_n) without any parallel data

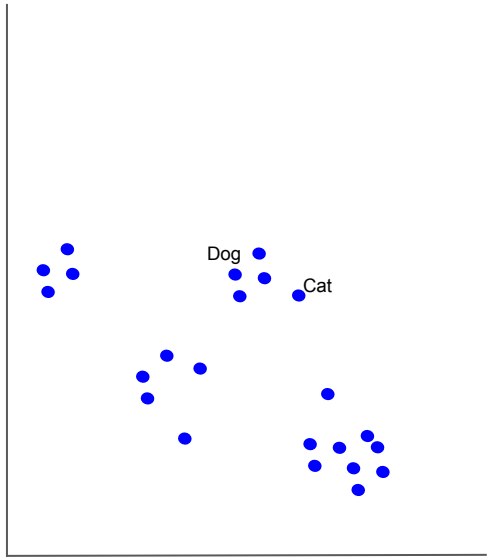
   We consider four (n=4) Languages: **English** (L_1), **French** (L_2), **German** (L_3), **Spanish** (L_4)

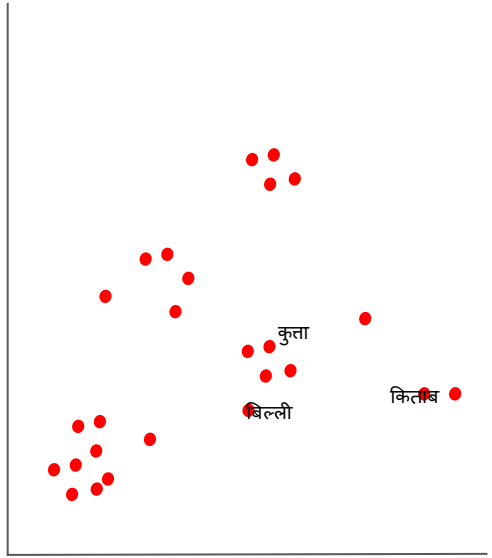
2. Map L_2, L_3, …, L_n into the Embedding space of L_1

   Map **French, German, Spanish** into the Embedding space of **English**

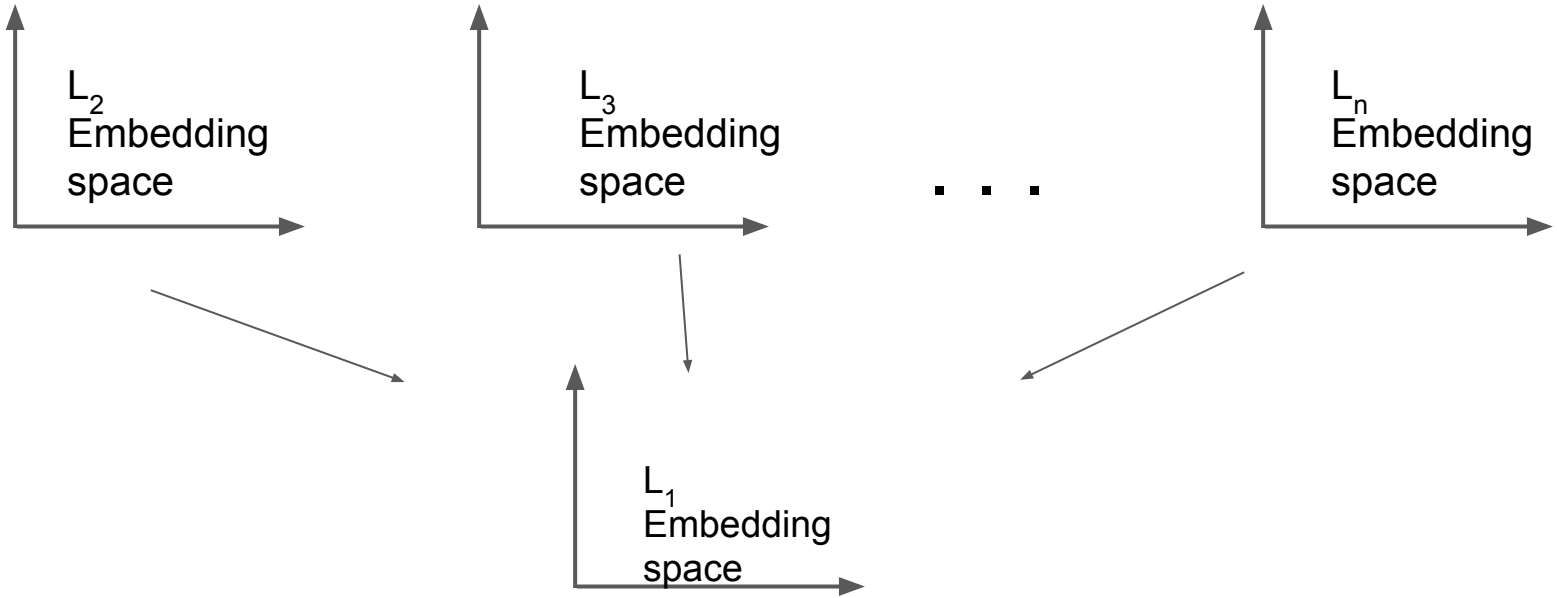
Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal and Pushpak Bhattacharyya. *Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders.* 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy. 2019

Before Mapping



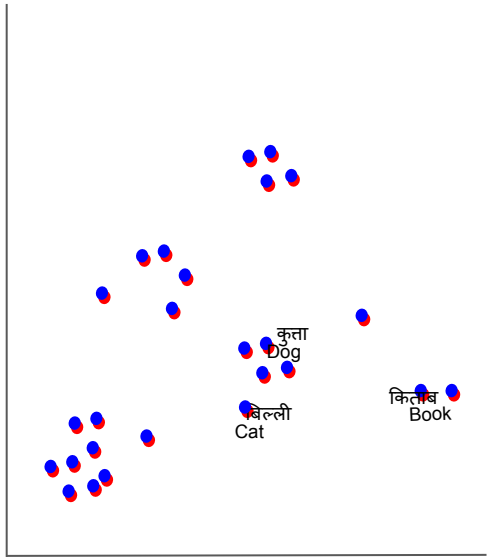


Multilingual Embedding



We pairwise map all non-English embedding spaces into the English embedding space using Conneau et al. (2018)

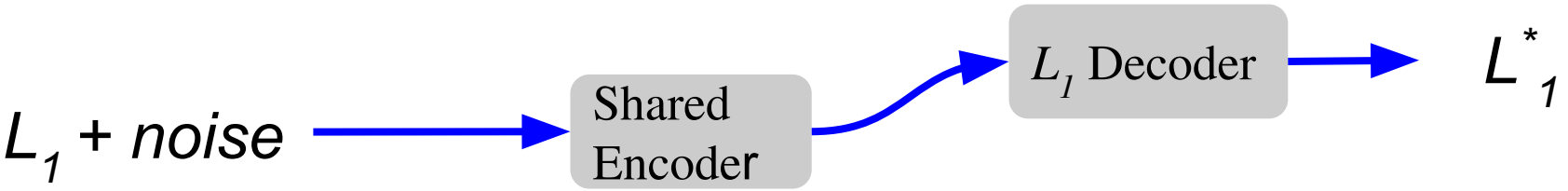
After Mapping



Our approach

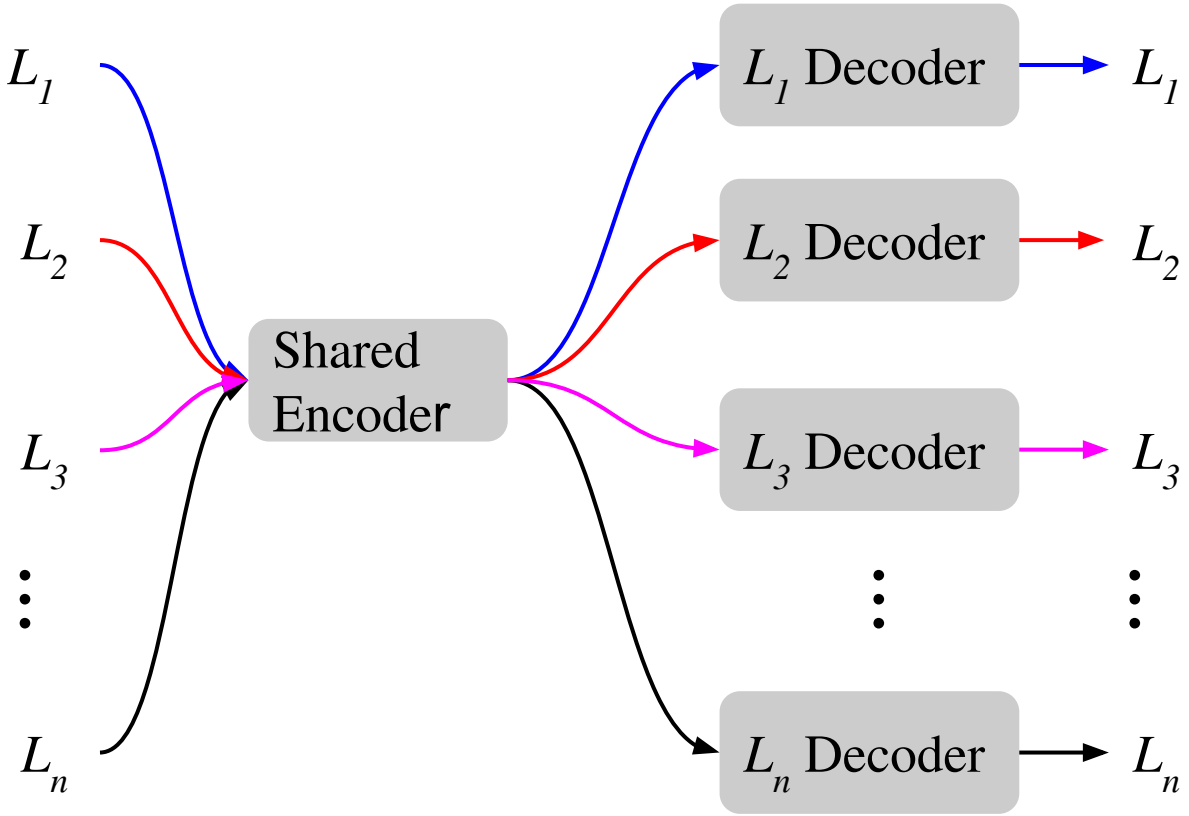
3. Denoise all languages (L_1, L_2, L_3, …, L_n)

Denoising for L_1



Loss between L^*_1 and L_1

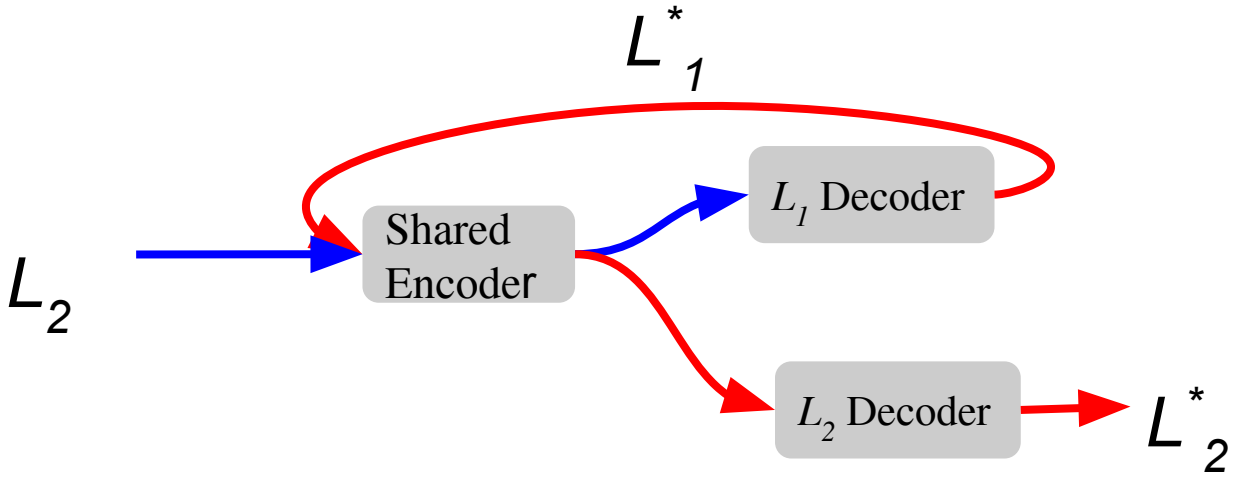
Denoising Autoencoding



Our approach

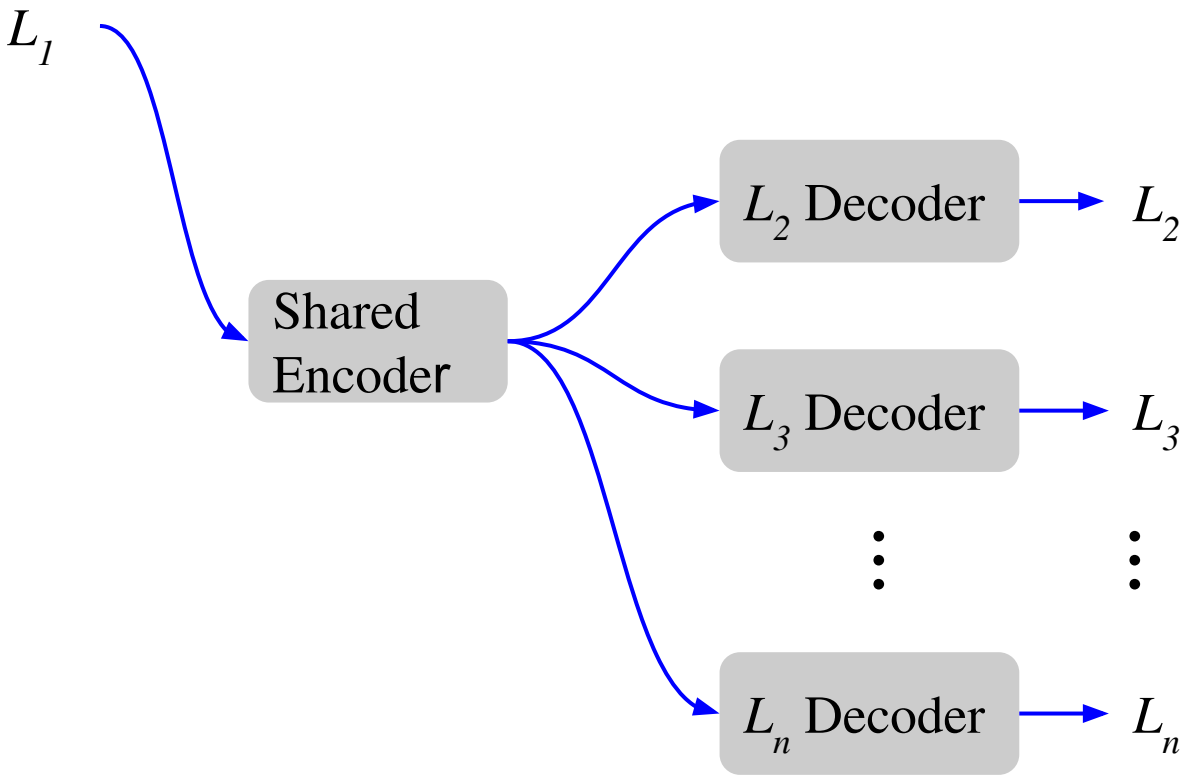
4. Back-translate between L_1 and (L_2, L_3, …, L_n) only
   - Back-translate between English and Non-English only
   - ***By-product****: Zero-shot translation -- network learns to translate between unseen language pairs too*

Back-translation step: $L_1 \rightarrow L_2$

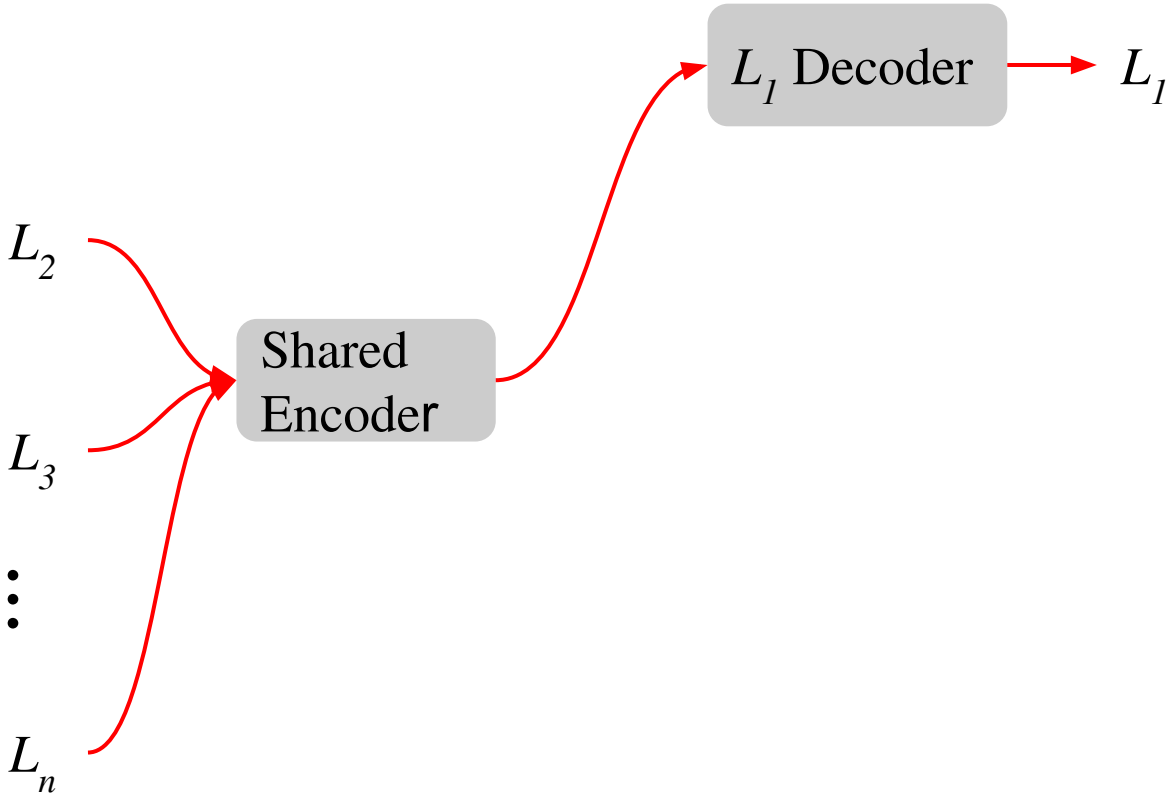


Loss between L^*_2 and L_2

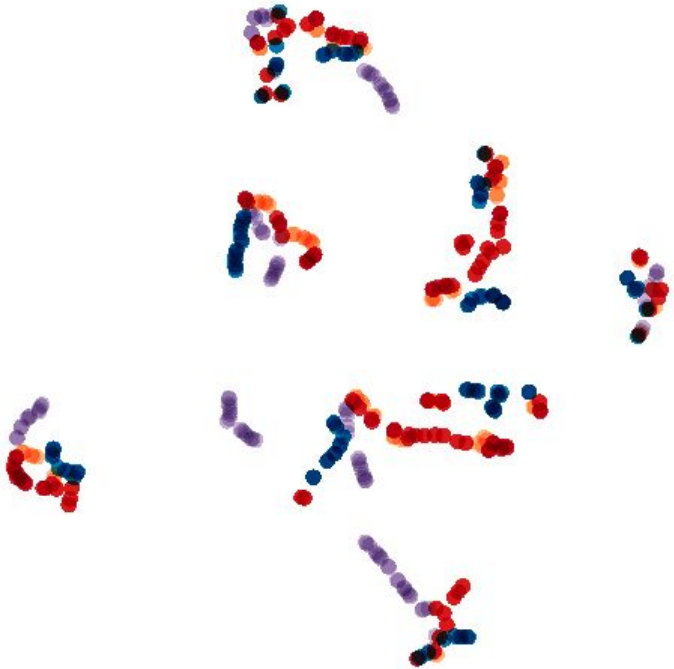
Back-translation (English to Non-English)



Back-translation (Non-English to English)



Benefit of Shared Encoder



- German
- English
- French
- Spanish

t-SNE projection of hidden vectors obtained from the shared encoder for some sentences in four languages. Each cluster indicates one sentence in four languages. Dots are the words in a sentence. Color represents the languages.

Results

System	*newstest2013*			*newstest2014*		
	Base	**Multi**	**Improvement**	**Base**	**Multi**	**Improvement**
Fr → En	13.81	14.47	+0.66	14.98	15.76	+0.78
Es → En	13.97	15.45	+1.48	-	-	-
En → Fr	13.28	13.71	+0.43	14.57	14.69	+0.12
En → Es	14.01	14.82	+0.81	-	-	-
De → En	11.30	11.94	+0.64	10.48	11.21	+0.73
En → De	7.24	8.09	+0.85	6.24	6.77	+0.53

- BLEU scores on *newstest2013* and *newstest2014.* Spanish (Es) is not part of *newstest2014* test set. **Base**: Baseline and **Multi**: Multilingual

Sample Outputs (French→ English)

Source	Reference	Bilingual	Multilingual
La préparation à gérer une classe dans un contexte nord-américain, québécois.	Preparation to manage a class in a North-American and Quebec context.	The build-up to manage a class in a Australian, Australian.	The preparation to handle a class in a Latin American context.
Il va y avoir du changement dans la façon dont nous payons ces taxes.	There is going to be a change in how we pay these taxes.	There will be the change in the course of whom we owe these bills.	There will be the change in the way we pay these taxes.

Sample Outputs (German→ English)

Source	Reference	Bilingual	Multilingual
Auch diese Frage soll letztlich Aufschluss darüber geben, welche Voraussetzungen es für die Entstehung von Leben gibt.	This question should also provide information regarding the preconditions for the origins of life.	This question will also ultimately give clues about what there are for the evolution of life.	This question will ultimately give clues to how there is conditions for the emergence of life.
Ihm werde weiterhin vorgeworfen, unerlaubt geheime Informationen weitergegeben zu haben.	He is still accused of passing on secret information without authorisation.	Him will continue to be accused of stealing unlawful information.	Him would continue to be accused of illegally of leaking secret information.

Sample Outputs (Spanish→ English)

Source	Reference	Bilingual	Multilingual
Los estudiantes, por su parte, aseguran que el curso es uno de los más interesantes.	Students, meanwhile, say the course is one of the most interesting around.	The students, by their part, say the practice is one of the most intriguing.	The students, by their part, say the course is one of the most interesting.
No duda en contestar que nunca aceptaría una solicitud de una persona desconocida.	He does not hesitate to reply that he would never accept a request from an unknown person.	No doubt ever answering doubt it would never accept an argument an unknown person.	No doubt in answer that he would never accept a request of a unknown person.

Results (Zero-shot Translation in Unsupervised NMT)

⟶	Es	Fr	De
Es	-	13.92	4.78
Fr	13.87	-	4.59
De	7.40	6.78	-

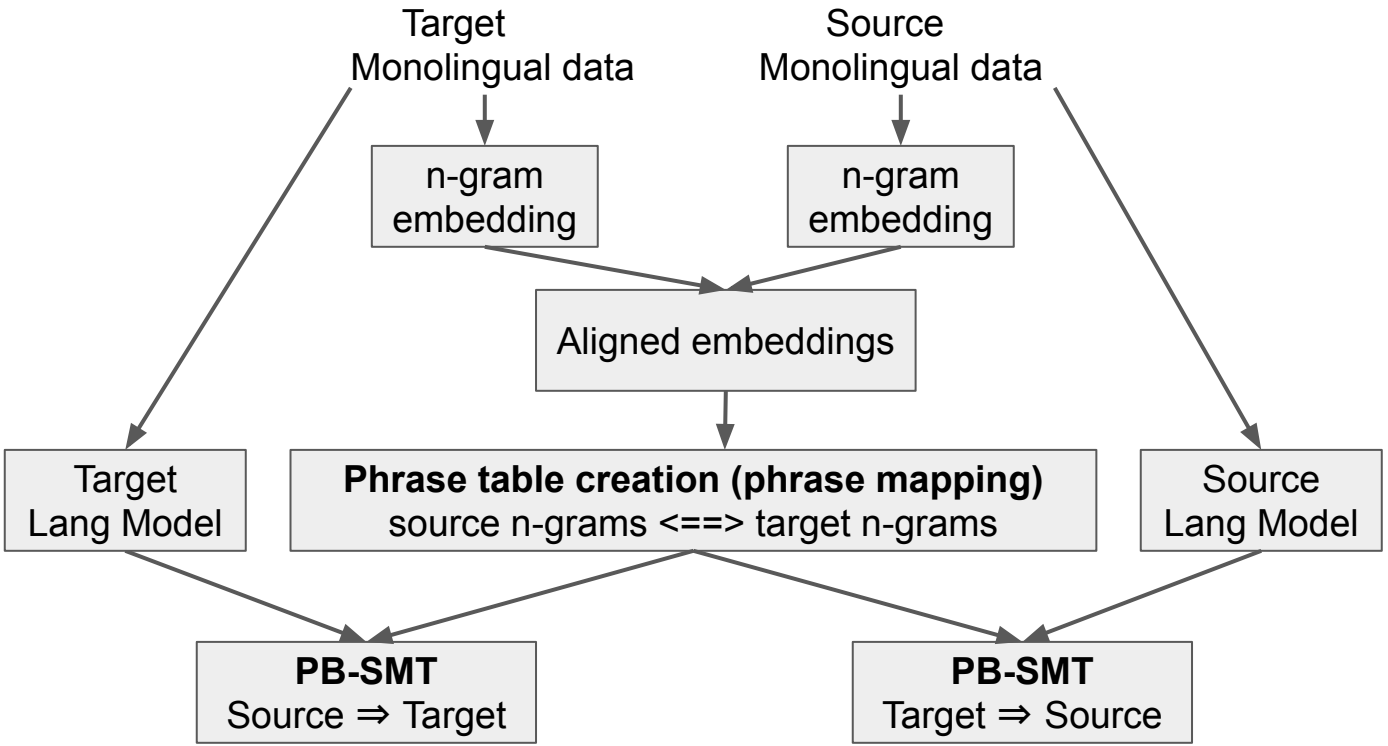
- BLEU scores of translation between non English languages on *newstest2013*. Consider rows are source and columns are target. The network is not trained for these language pairs and still it is possible to translate between these pairs by using the shared encoder and language specific decoders

Phrase-based & Neural Unsupervised Machine Translation

- **PB-SMT requires**
  - Phrase Table (needs parallel sentences)
  - Target-side Language Model (needs only monolingual sentences)
- **Unsupervised PB-SMT requires**
  - Phrase Table (***will be learned using n-gram source-target embedding mapping***)
  - Target-side Language Model (***needs only monolingual sentences***)
- **Unsupervised (PB-SMT + NMT)**
  - Add the data generated by the unsupervised PBSMT system to the back-translated data produced by the NMT model

***Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer and Marc'Aurelio Ranzato. Phrase-Based & Neural Unsupervised Machine Translation. *EMNLP* (2018)**

Phrase-based Unsupervised Machine Translation



Iterate over SMT model using the monolingual data multiple times and each time the phrase table gets improved

Phrase-based & Neural Unsupervised Machine Translation: Results

- **Training data size:**

English	German	French	Russian	Romanian
495m	622m	122m	12m	2.9m

- **BLEU Score:**

	en → fr	fr→ en	en→ de	de→ en	en→ ro	ro→ en	en→ ru	ru→ en
Unsupervised PBSMT								
Unsupervised phrase table	-	17.50	-	15.63	-	14.10	-	8.08
Back-translation - Iter. 1	24.79	26.16	15.92	22.43	18.21	21.49	11.04	15.16
Back-translation - Iter. 2	27.32	26.80	17.65	22.85	20.61	22.52	12.87	16.42
Back-translation - Iter. 3	27.77	26.93	17.94	22.87	21.18	22.99	13.13	16.52
Back-translation - Iter. 4	27.84	27.20	17.77	22.68	21.33	23.01	13.37	**16.62**
Back-translation - Iter. 5	**28.11**	27.16	-	-	-	-	-	-

MT in Indian languages

(RBMT, EBMT, SMT, NMT & Hybrid)

Rule-based MT in Indian languages

AnglaHindi:An English to Hindi Machine-Aided Translation System (1/2)

- **Rule-based Machine Aided Translation** = Interlingua + Example-based translation methodology
- **AnglaHindi** > Derived from Anglabharati system which translates English to Indian languages
- Intermediate structure created after the disambiguation
- Intermediate language structure has word and word-group order per the structure of the group of target languages
- Intermediate structure converted to each Indian language through a process of text-generation

Sinha, R. M. K., and A. Jain. AnglaHindi: an English to Hindi machine-aided translation system. *MT Summit IX, New Orleans, USA* 494 (2003): 497.

Anusaaraka: An Expert System based Machine Translation System

- Morphological analyser
- Part-of-speech tagger
- **Lexical transfer module**: With the help of bilingual dictionary, source language lexical form will be converted to target language lexical form
- **Morphological generator**: Generates target language surface form for each target language lexical form
- **A structural transfer module**: A finite-state chunker to detect the patterns of lexical forms which needs to be processed for word reordering, agreement etc. and performs those operations

Chaudhury, Sriram, Ankitha Rao, and Dipti M. Sharma. "Anusaaraka: An expert system based machine translation system." ***Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*****. IEEE, 2010.**

A Domain-Restricted, Rule based, English-Hindi Machine Translation System based on Dependency Parsing

- Transfer phase is replaced with a syntax planning algorithm
- Transfer and Generation phases combined together into single generation phase
- The advantage with this approach is that it directly linearizes the dependency parse of the source sentence as per the syntax of target sentence
- **Word translation:** Target word for each source word is looked up in domain-specific dictionary
- **Feature transfer**
  - Nouns to adjectives: Adjectives need to take the number and gender information from the noun that it qualifies
  - Nouns to verbs: For verbs, gender and number information are obtained from the subject in case of active voice and object in the case of passive voice
- **BLEU score**: 18-27

Desai, Pratik, Amit Sangodkar, and Om P. Damani. "A domain-restricted, rule based, English-Hindi machine translation system based on dependency parsing. *Proceedings of the 11th International Conference on Natural Language Processing*. 2014.

Example based MT in Indian languages

A Phrasal EBMT System for Translating English to Bengali (contd..)

- **Input:** “A man had given the boy a good book in the school.”
- S**hallow parser** identifies the phrases as:
  - “<np1> A man </np1> <vp1> had given</vp1> <np2> the boy </np2> <np3> a good book </np3> <pp1> in the school </pp1>.”
- Phrase translator module translates the phrases as:
  - ‘A man’ **<===>** ‘ekjon lok’, ‘had given’ **<===>** ‘diyechhilo’, ‘the boy’ **<===>** ‘chheletike’, ‘a good book’ **<===>** ‘ekti bhalo boi’ and ‘in the school’ **<===>** ‘bidyalaye’.
- Re-order translated Bengali phrases using heuristics that are based on phrase ordering rules of Bengali
- **Output:** “<ekjon lok>^{1} <chheletike>^{3} <bidyalaye>^{5} <ekti bhalo boi>^{4} < diyechhilo >^{2}”.

***Naskar, S.K. and Bandyopadhyay, S., 2006, October. A phrasal EBMT system for translating English to Bengali. In *Satellite Workshop* (p. 69).**

A Hybrid Approach to Example based Machine Translation for Indian Languages

- **Hybrid EBMT**= Statistical MT + Minimal linguistic resources
- EBMT performs three distinct phases in order to transform a new input sentence into target language
  - **Matching phase**: Search the source side of parallel corpus for closest matches and their translations
  - **Alignment phase:** Determine the sub-sentential translation links in those retrieved examples
  - **Recombination phase**: Recombine the relevant parts of target translation links to derive the translation
- Using the example database, every input sentence can be segmented into the longest possible available fragments
- Alignments are first obtained from GIZA++ and then further enhanced by manual dictionary

Ambati, Vamshi, and U. Rohini. A hybrid approach to example based machine translation for Indian languages." *ICON* 2007 (2007)

A Hybrid Approach to Example based Machine Translation for Indian Languages (contd..)

- Since the phrases obtained are already aligned to target sentence, a simple combination will give good translation
- Training: 53k English-Hindi parallel sentences; Test set: 100 sentences

Method	BLEU
Word-word matching (manual dictionary)	12.4
Word-word matching (manual and statistical dictionary)	21.4
Proposed approach	43.2

Statistical based MT in Indian languages

Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT

- A factored based SMT model for English-to-Hindi translation
- **Factored based machine translation:**
  - Works on **more general representations, such as lemmas** instead of surface forms of words
  - Can draw on richer statistics and overcome the data sparseness problems caused by the limited training data
  - Best explain on a morphological, syntactic, or semantic Level
  - Having such information available to the translation model allows the direct modeling of these aspects

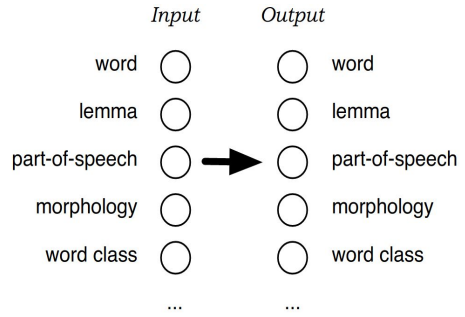


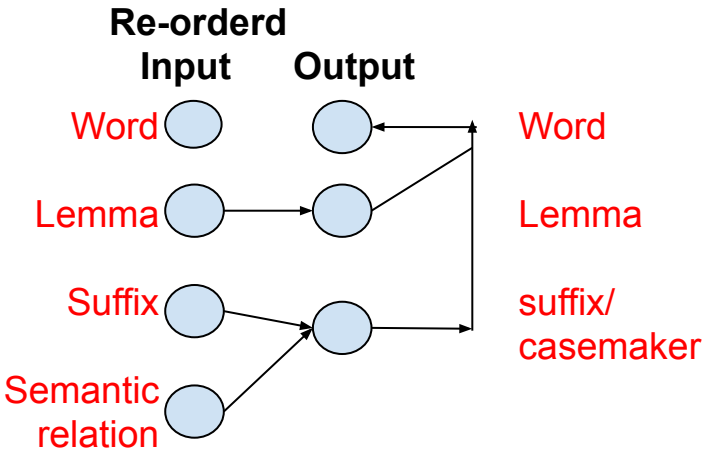
Figure 1: Factored representations of input and output words incorporate additional annotation into the statistical translation model.

- In PB-SMT, any instance of ***'house'*** doesn't hold information of ***'houses'***
- In factored SMT, if model knows ***'house'*** it can relate to ***'houses'*** since lemma of ***'houses'*** is ***'house'***

*Koehn, P. and Hoang, H., 2007, June. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 868-876).

Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT

- Hindi-side factors: Case markers and suffixes
- English-side factors: Semantic relation and suffixes
- Hindi is a relatively free word-order language
- Constituents (sub., obj. etc) can move around in the sentence without impacting the core meaning
  - जॉन ने मेरी को देखा *(John ne Mary ko dekhaa)*
  - मेरी को जॉन ने देखा (*Mary ko John ne dekhaa)*

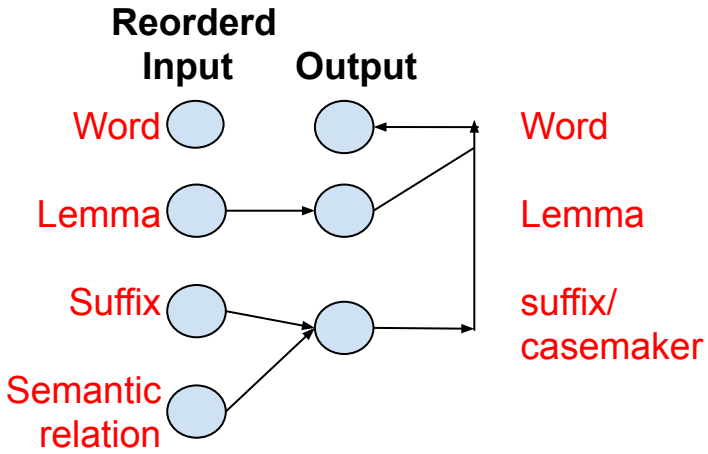


- Both the sentences are similar. The identity of ***John*** as the subject and ***Mary*** as the object in both sentences comes from the case markers ***ने*** (ne– nominative) and ***को*** (ko–accusative). Therefore, even though Hindi is predominantly SOV in its word-order, *correct casemarking is a crucial part of making translations convey the right meaning*

***Ramanathan, A., Choudhary, H., Ghosh, A. and Bhattacharyya, P., 2009, August. Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 800-808).**

Case markers and morphology:Addressing the crux of the fluency problem in English-Hindi SMT

- **Semantic relation**: how two entities in a sentence are related?
- Target language suffixes are largely determined by source language suffixes and casemarkers (which in turn are determined by the semantic relations)
  - *The boys ate apples.*
  - *The|empty|det boy|s|subj eat|ed|empty apple|s|obj*
  - लड़को ने सेब खाये *(ladakon ne seba khaaye)*
  - Here, the plural suffix on ***'boys'*** leads to two possibilities



'लड़के' (ladake– plural direct) and '***लड़को'*** (ladakon– plural oblique). The case marker '***ने'*** (ne) requires the oblique case

- For semantic relations, Stanford parser and Universal semantic relations are used

***Ramanathan, A., Choudhary, H., Ghosh, A. and Bhattacharyya, P., (2009). Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 800-808).**

Datasets and Evaluation Results

- **Training data size:** 12,868 parallel English-Hindi sentences

- **Results**

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

Neural MT in Indian languages

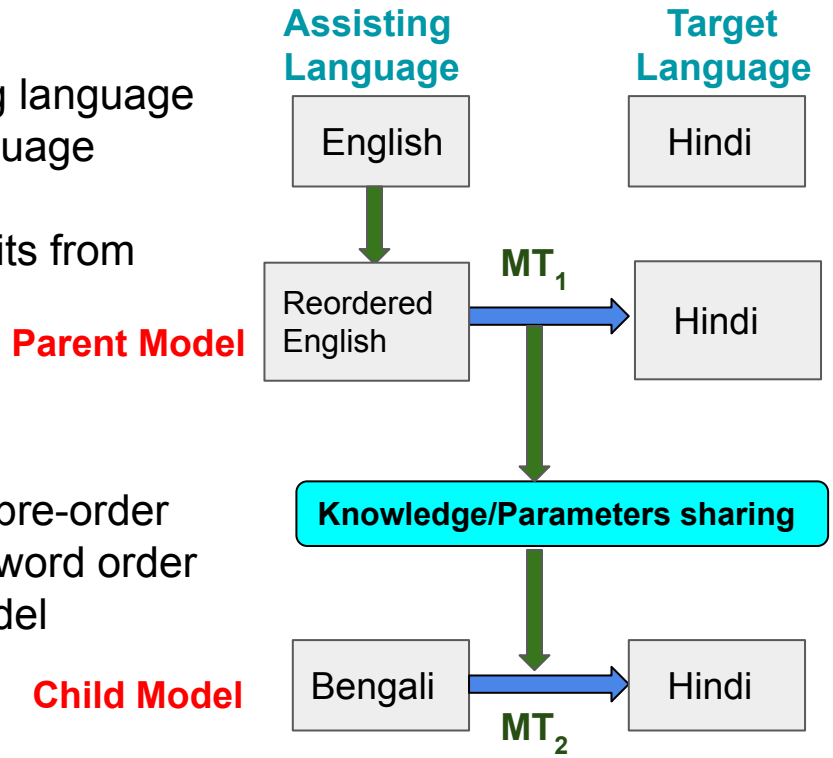
Transformer-based Multilingual Neural Machine Translation System

- Multilingual Neural Machine Translation
  - English and 7 Indic languages
  - Many-to-one and One-to-many directions
- Multilingual models: additional token indicating which Indic language a sentence pair belongs to is added at the beginning of every source sentence. e.g.
  - HI## I need your help. मुझे आपकी मदद चाहिए ।
  - BN## Where the mind is without fear! চিত্ত যেথা ভয় শূন্য!
- Combine all training data to train a single model to translate between all languages
- Language pair: English -- [Hindi, Bengali, Malayalam, Tamil, Telugu, Urdu, Sinhalese]

S. Sen, K. Gupta, A. Ekbal and P. Bhattacharyya (2018). IITP-MT at WAT 2018: Transformer-based Multilingual Indic-English Neural Machine Translation System, WAT, PACLIC

Addressing Word-order Divergence in Multilingual Neural Machine Translation for Extremely Low Resource Languages (1/3)

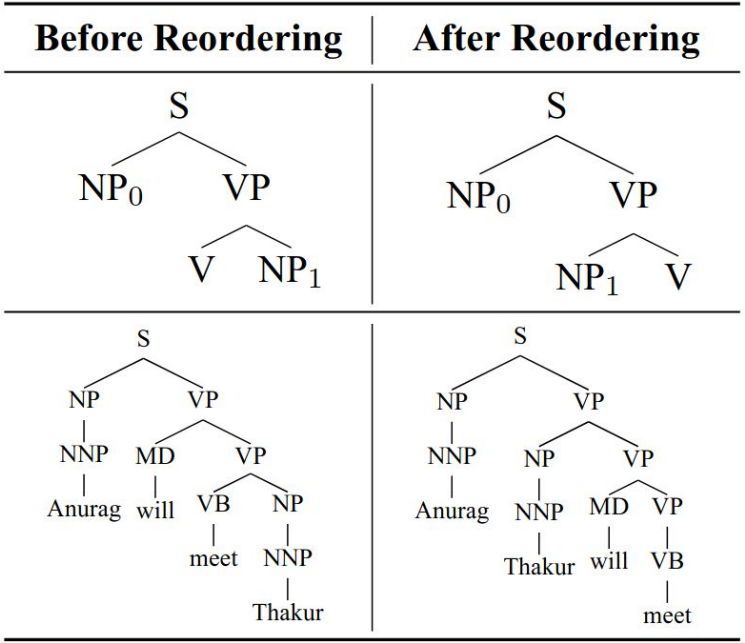
- In many cases of transfer learning, the assisting language has a different word order from the source language
- Divergent word order adversely limits the benefits from transfer learning
  - English: SVO
  - Hindi: SOV
- To bridge this divergence, authors proposed to pre-order the assisting language sentences to match the word order of the source language and train the parent model



*Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages. In NAACL

Addressing Word-order Divergence in Multilingual Neural Machine Translation for Extremely Low Resource Languages (2/3)

- Reordered parent model (English-Hindi) will be trained and used to initialize the child model using transfer learning
- A rule based approach is used to reorder English sentence (SOV) to Hindi (SVO)
- Example:
  - Original: Anurag will meet Thakur
  - Reordered: Anurag Thakur will meet
  - Hindi: अनुराग ठाकुर से मिलेंगे
  - anuraag thaakur se milenge



Addressing Word-order Divergence in Multilingual Neural Machine Translation for Extremely Low Resource Languages (3/3)

- **Parent model:** English-Hindi
- **Child models:**
  - Gujarati-Hindi, Marathi-Hindi, Tamil-Hindi, Bengali-Hindi, Malayalam-Hindi
- **Bleu Scores:**

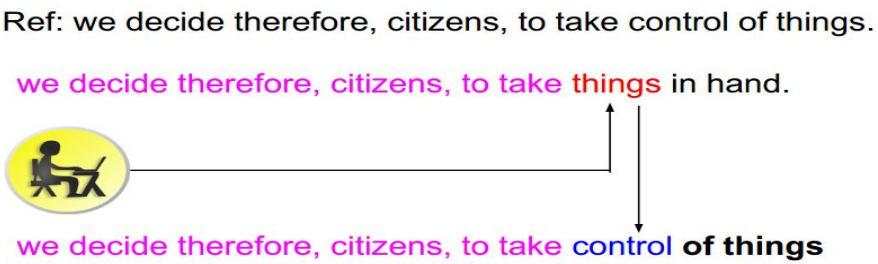
	No pre-order	**Pre-order**
Gujarati-Hindi	9.81	**14.34**
Marathi-Hindi	8.77	**10.30**
Tamil-Hindi	4.86	**6.00**
Bengali-Hindi	6.72	**9.19**
Malayalam-Hindi	5.73	**6.95**

Training data size

English-Hindi	1.5m
Gujarati-Hindi	49,999
Marathi-Hindi	49,999
Tamil-Hindi	49,999
Bengali-Hindi	49,999
Malayalam-Hindi	49,999

Syntax-informed Interactive Neural Machine Translation

- **Why Interactive?**
  - Human translators correct errors obtained from an automatic translation system in collaboration with the MT systems
  - Users may insert/replace their choice of words in the hypothesis generated by the model
  - Assists in ***Post-editing***



K. K. Gupta, R. Haque, A. Ekbal, P. Bhattacharyya, A. Way (2020). Syntax-informed interactive neural machine translation. IJCNN 2020

Syntax-informed Interactive Neural Machine Translation

- Model regenerates a new hypothesis which preserves the users' choice of words and contextually depends on the inserted token
- Multiple attempts of token replacements may be required by a user to obtain the desired output

- External Syntactic information: **CCG Supertags**
  - Known to be context sensitive tags that preserve the global syntactic information at local lexical level
  - Having this property, supertags resolve ambiguity in short- and long-distance dependencies by capturing the previous and next syntactic dependencies of a lexical term

K. K. Gupta, R. Haque, A. Ekbal, P. Bhattacharyya, A. Way (2020). Syntax-informed interactive neural machine translation. IJCNN 2020

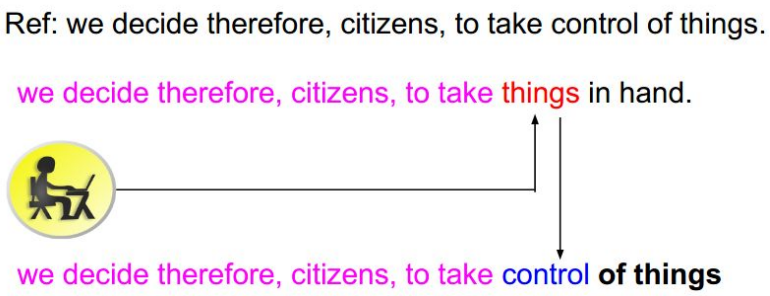
Syntax-informed Interactive Neural Machine Translation

In NMT:

- Output hidden state: $S_{i+1} = g(S_i, Y_i, C_i)$
- Output token: $Y_{i+1} = f(S_{i+1}, Y_i)$

In INMT:

- User makes change at the position T_i.
  - **Y_i is changed to y_i**
    - **Y_i is decoder generated token**
    - **y_i is user modified token**
- Output hidden state: $S_{i+1} = g(S_i, y_i, C_i)$
- Output token: $Y_{i+1} = f(S_{i+1}, y_i)$
- Now, the next token at position ***i+1*** will be generated based on y_i
- ***Using CCG supertags at target side signifies what exactly a particular lexical term is expecting in the sequence.***



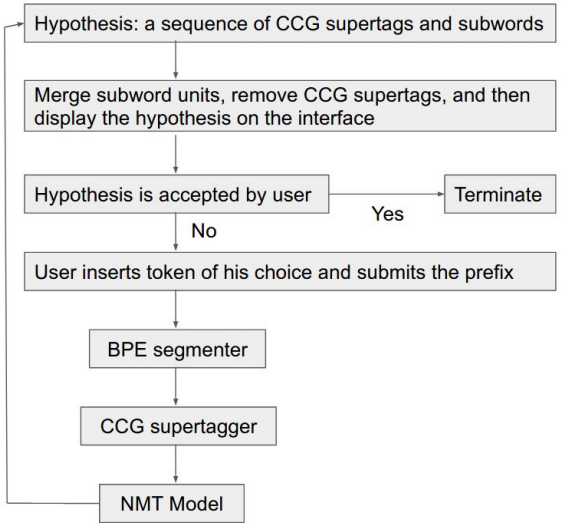
- In the given figure, ***'things'*** is changed to ***'control'*** by user.
- New token ***'of'*** is generated after user's token ***'control'***.

Syntax-informed Interactive Neural Machine Translation

- **Objective:** To reduce the human efforts (token replacement) in translation in an interactive-predictive platform by increasing the prediction accuracy of the decoder
- **Language Pairs:** Hindi-English and French-English
- **Evaluation metrics:**
  - *WPA:* Word prediction ratio
    - Ratio of total correct predicted tokens
  - *WSR:* Word stroke ratio
    - Ratio of total replaced tokens

Language Pair	#sentences
Hindi-English	1.5m
French-English	12m

		Baseline	On-the-fly CCG supertagger
Fr → En	WPA	46.82	**49.47 (2.65 ↑)**
	WSR	53.77	**50.61 (3.16 ↓)**
Hi → En	WPA	34.32	**40.87 (6.55 ↑)**
	WSR	65.68	**59.12 (6.56 ↓)**



Syntax-informed Interactive Neural Machine Translation

1	Input sentence (bpe)	jamais joué à un jeu à b@@ oire basé sur le thème N@@ azi cela dit .
2	Reference	never played a Nazi themed drinking game though .
3	Initial Hypothesis	never played a Nazi drinking play there .
4	Hypothesis after several iterations	NP never S[pss]\NP played NP/NP a N/N Nazi N them@@ N ed **(S[dcl]\NP)/NP play** (NP\NP)/NP though N .
5	INMT interface	never played a Nazi themed **play** though .
6	Correction by user	never played a Nazi themed **drinking** though .
7	Applying on-the-fly CCG supertagger	NP never S[pss]\NP played NP/NP a N/N Nazi N them@@ N ed **N drinking** (S\NP)\(S\NP) though N/N .
8	New hypothesis	never played a Nazi themed **drinking game** though .

TABLE: An example showing applying On-the-fly CCG supertagger on hypothesis. As can be seen from rows 5 and 6, the user replaces incorrect token ***play*** with correct token ***drinking***. The new token ***drinking*** gets the CCG supertag of the incorrect token ***play***, ***(S[dcl]\NP)/NP***, which is also incorrect. In second setup, On-the-fly CCG supertagger is applied on hypothesis (validated prefix and suffix). As can be seen from row 7, a new CCG supertag sequence is generated for the hypothesis, and we see that the CCG supertag ***(N)*** is assigned to the new token ***drinking***. In row 8, new hypothesis is generated with ***game*** as the next predicted token.

Modelling Source- and Target-Language Syntactic Information as Conditional Context in INMT

- CCG supertags at the target side + Syntactic features from constituency parsers at source side
- For using information from constituency parsers at the source side, *we extract a chunk sequence from the constituency parse tree of a source sentence by setting random a maximum chunk size ({1...6}) for every sentence*
- Here the tags are in the form of:
  - ***Chunk_identifier+Chunk size***
  - **For example:** ***PP2 de 2008***, here PP is chunk identifier and 2 in chunk size i.e. it is covering two tokens *"de 2008".*
  - **Source:** il y a des voitures neuve et ch`ere`a tout les coins de rue, exactement comme avant la crise de 2008.
  - **Source with chunk info: VN3** il y a **DET1** des **NC1** voitures **AP3** neuve et ch`ere **P1**`a **ADJ1** tout **DET1** les **NC1** coins **P1** de **NC1** rue **PONCT1** , **ADV1** exactement **P1** comme **P1** avant **DET1** la **NC1** crise **PP2** de 2008 **PONCT1** .

K. K. Gupta, R. Haque, A. Ekbal, P. Bhattacharyya, A. Way (2020). Modelling Source- and Target-Language Syntactic Information as Conditional Context in Interactive Neural Machine Translation. EAMT 2020

Results

- **Language Pairs:** French-English
- **Evaluation metrics:**
  - *WPA:* Word prediction ratio
    - Ratio of total correct predicted tokens
  - *WSR:* Word stroke ratio
    - Ratio of total replaced tokens

Language Pair	#sentences
French-English	12m

		Baseline	On-the-fly CCG supertagger
Fr → En	WPA	46.82	**49.47 (2.65 ↑)**
	WSR	53.77	**50.61 (3.16 ↓)**

Syntax info at Target side only

		Baseline	On-the-fly CCG supertagger
Fr → En	WPA	46.82	**51.12 (4.30 ↑)**
	WSR	53.77	**48.93 (4.84 ↓)**

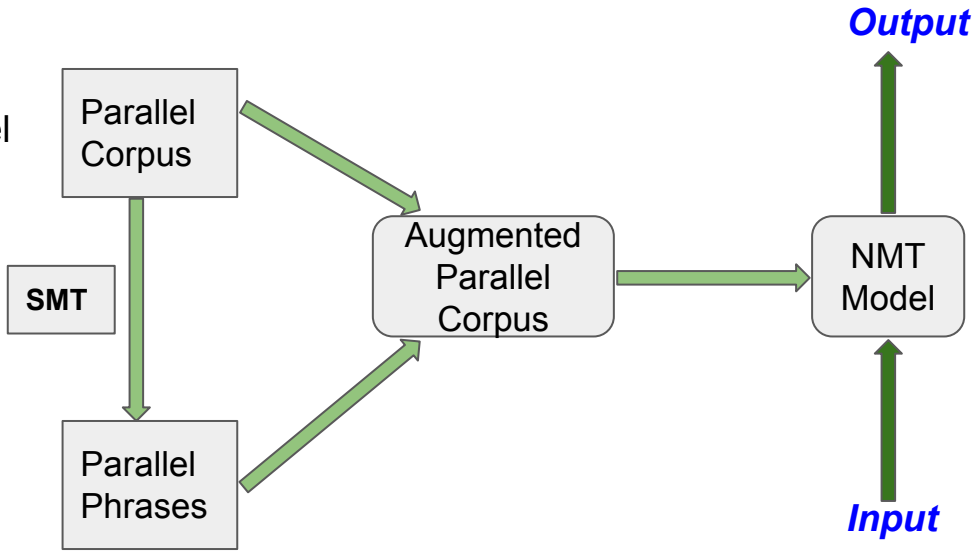
Syntax info at target and source side both

- ***We can see that source and target-side syntactic contextual features complement each other as far as neural interactive prediction is concerned.***

Hybrid MT in Indian languages

Neural Machine Translation of Low-resource Languages using SMT Phrase Pair Injection

- An SMT-NMT hybrid system
- Augment the original parallel sentences with parallel phrases from the phrase table
- Use augmented parallel corpus to train the NMT model
- Parallel phrases enrich the training data to help low-resource NMT model



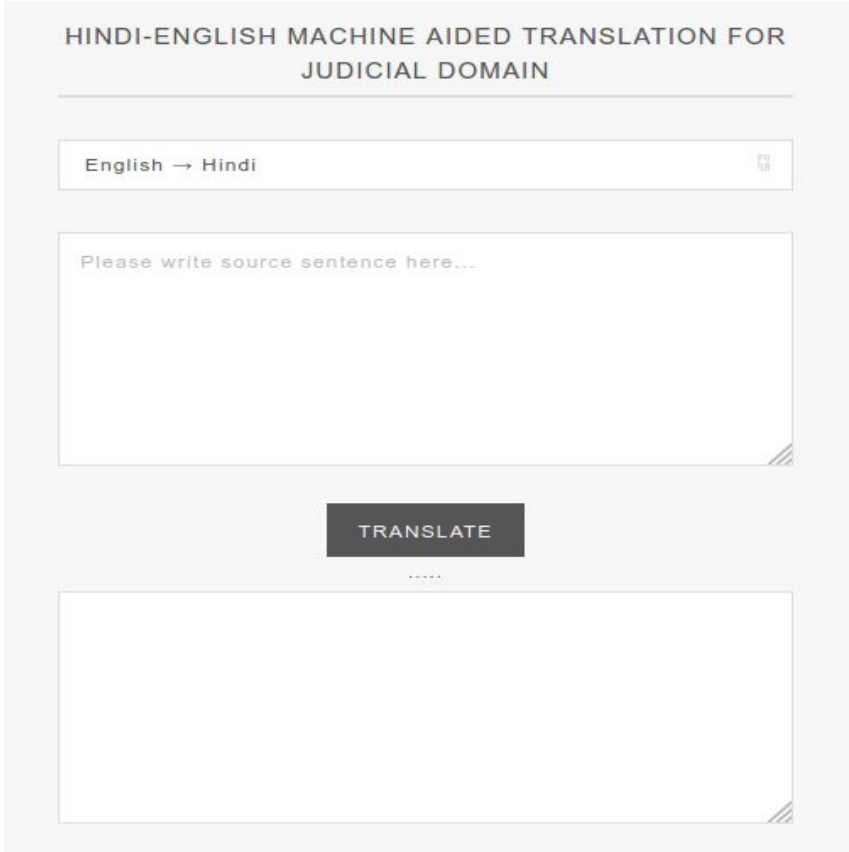
*Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. Neural machine translation of low-resource languages using SMT phrase pair injection. Natural Language Engineering, Cambridge University Press, pages 1–22, June 2020

MT Services in Indian Languages

Hindi-English Machine Aided Translation System (HEMAT)

- A web-based Hindi-English machine aided translation system for judicial domain
- **Domain** : Judicial
- A consortium leaded by IIT Patna
- **Technique** : NMT with self and multihead attention and Syntactic Phrase Augmentation
- **Data**: Collected from various government websites (RTI, PM India etc.) and judicial domain corpus (court judgements, proceedings, part of legal reports etc.)
  - Total: 1,098,940 parallel sentences (**judicial**: approx. 3.5 Lacs)
- **Evaluation** :
  - Hindi→English : 55.47
  - English→Hindi : 56.09
- **Link** : https://www.iitp.ac.in/~hemat/

HEMAT: Interface



ANGLABHARATI

- First MT system in India
- Developed by a consortium of institutions including IIT Kanpur, IIT Bombay, IIT Guwahati, CDAC Kolkata,, CDAG-GIST group Pune, CDAC Thiruvananthpuram, TIET Patiala, JNU Delhi and Utkal University, Orissa in 1991
- Primarily from English to Hindi
- Later extended to all Indian languages in 2004, namely Assamese, Bengali, Urdu, Marathi, Konkani, Punjabi, Sanskrit, Oriya, Sindhi, Kashmiri
- **Approach**: Pseudo-interlingua
- **Domain**: General
- **Link**: http://tdil-dc.in/index.php?option=com_content&view=article&id=61&lang=en

ANUBHARATI

- Developed by IIT Kanpur
- Initial version released in 1995
- Primarily from Hindi to English
- Later extended to Hindi to any other Indian languages in 2004
- **Approach**: Generalized EBMT
- Domain: General

Anusaaraka

- Developed by IIT Kanpur, University of Hyderabad and IIIT Hyderabad
- Initial version was released in 1995
- Languages: Punjabi, Bengali, Telugu, Kannada, Marathi and English to Hindi
- **Approach**: Principle based on Paninian Grammar
- **Domain**: General

SHIVA and SHAKTI

- Both of the systems were developed by IISc Bangalore along with IIT Hyderabad, Carnegie Mellon University
- Initial version released in 2004
- SHIVA:
  - English to Hindi
  - Approach: EBMT
  - Domain: General
- SHAKTI:
  - English to Hindi, Marathi and Telugu
  - Approach: Hybrid approach by combining RBMT and SMT
  - Domain: General

Sampark

- Developed by a consortium of institutions including IIIT Hyderabad, University of Hyderabad, CDAC (Noida,Pune), Anna University, KBC Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore,IIIT Allahabad, Tamil University, Jadavpur University
- Consortium led by IIIT Hyderabad
- Supports 18 language pairs between Indian languages namely, Punjabi – Hindi –Punjabi, Telugu – Tamil – Telugu, Urdu – Hindi – Urdu, Hindi – Telugu – Hindi, Marathi – Hindi – Marathi, Bengali – Hindi – Bengali, Tamil – Hindi – Tamil, Kannada – Hindi – Kannada, Malayalam – Tamil – Malayalam
- Approach: Hybrid approach by combining RBMT and SMT
- Domain: General
- Link: http://tdil-dc.in/mt/common.php

Anuvadaksh

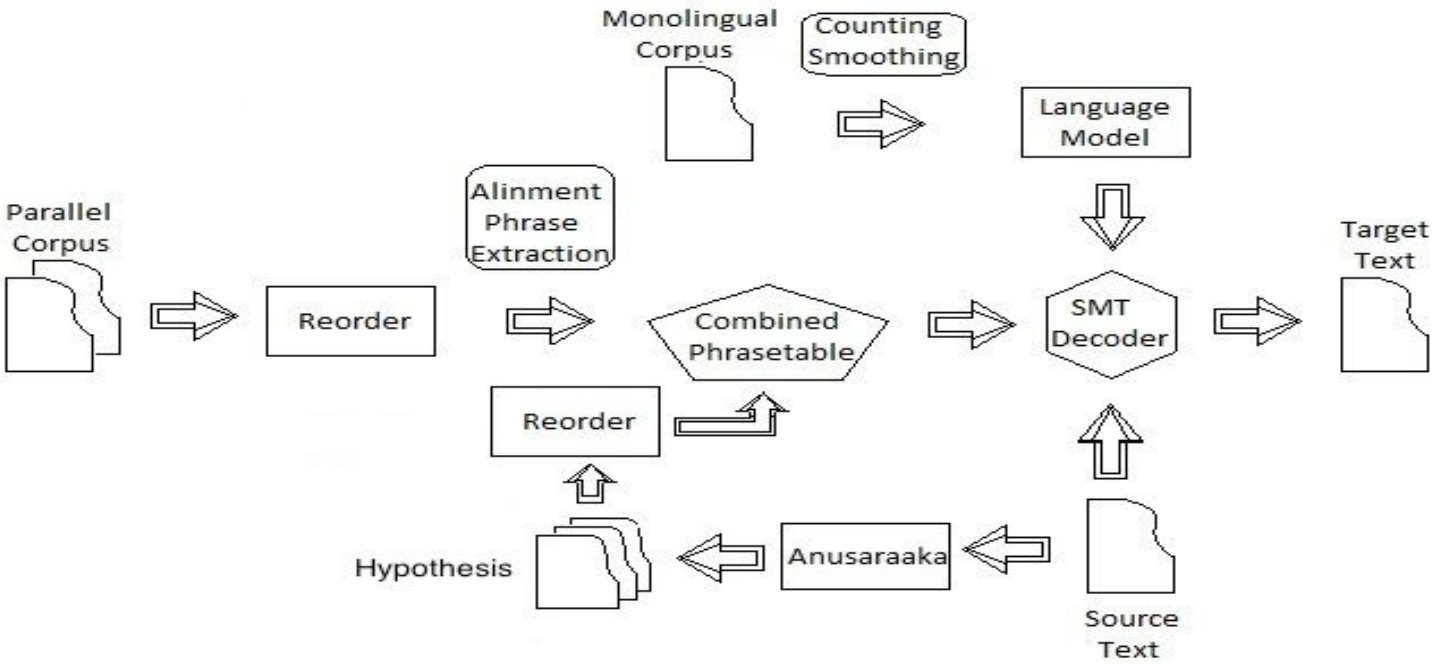
- Developed by English to Indian Language MT(EILMT) consortium
- Participating institutions: IIT Kharagpur, IIIT Hyderabad, University of Hyderabad, IIT Bombay, IIIT Kerala, AU-KBC Chennai, CDAC Noida
- English to Hindi, Urdu, Oriya, Bangla, Marathi, and Tamil (6 Indian languages)
- **Approach**: Hybrid approach with Tree-Adjoining-Grammar (TAG) based MT, SMT, EBMT and RBMT
- **Domain**: Tourism

Tools and Data for MT in Indian languages

- **Indic NLP Library**: A unified approach to NLP for Indian languages by Anoop Kunchukuttan
  - For processing involving Indic scripts/languages
  - Normalization, Tokenization/De-tokenization, Orthographic subword, many more
  - https://github.com/anoopkunchukuttan/indic_nlp_library
- ILCI Corpus between multiple Indic languages and English by Jha (2010)
- IIT Bombay English-Hindi Corpus by Kunchukuttan et al. (2018)
  - (Download: http://www.cfilt.iitb.ac.in/iitb_parallel/)
- A Catalog of resources for Indian language NLP by Anoop
  - Pointer to everything related to Indian languages https://github.com/indicnlpweb/indicnlp_catalog

Our Research at IIT Patna

Serial Coupling of SMT and RBMT



Debajyoty Banik, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. Can SMT and RBMT Improve each other's Performance? - An Experiment with English-Hindi Translation. International Conference on Natural Language Processing (ICON 2014). 2016.

Serial Coupling of SMT and RBMT: Dataset

English-Hindi parallel corpus sentences from product domain

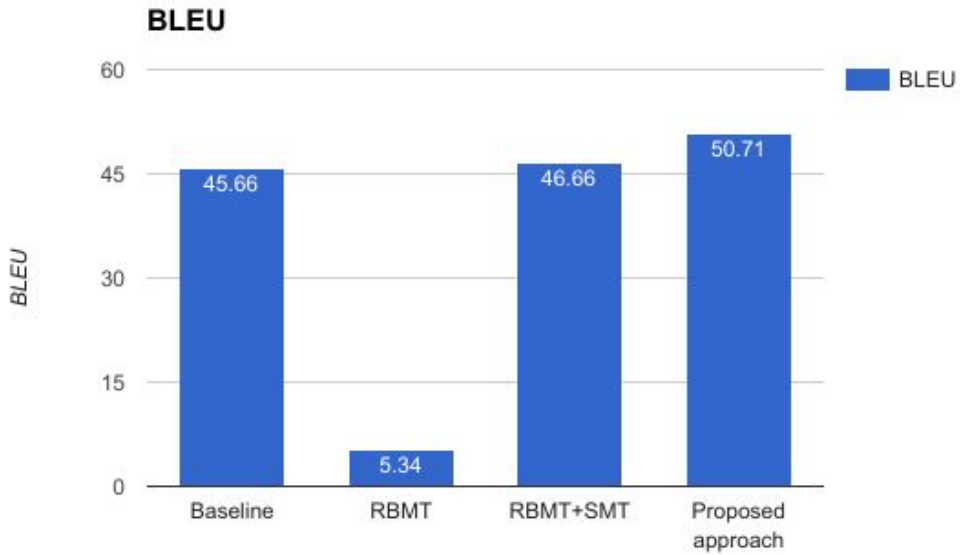
Example

- To know more about app click here
- ऐप के बारे में अधिक जानकारी के लिये यहां क्लिक करें

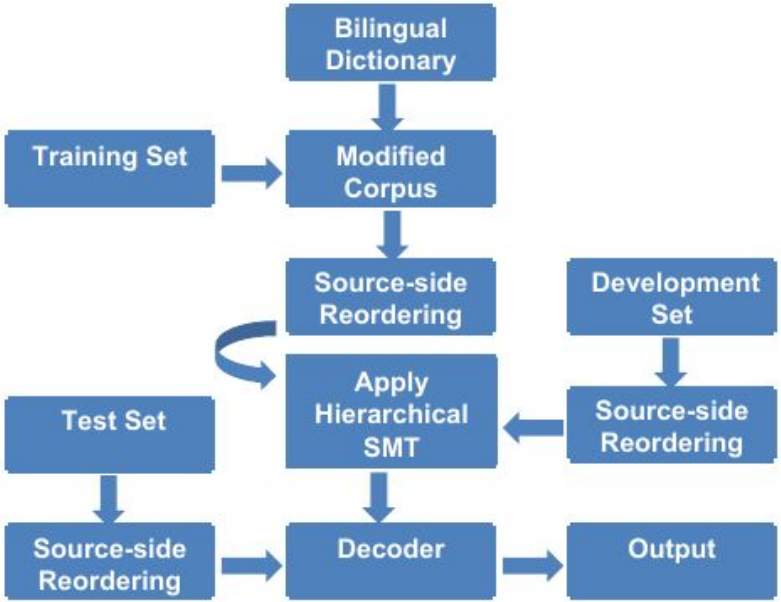
Set	#Sentence
Train	111,586
Development	602
Test	5640

Debajyoty Banik, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. Can SMT and RBMT Improve each other's Performance? - An Experiment with English-Hindi Translation. International Conference on Natural Language Processing. 2016.

Serial Coupling of SMT and RBMT: Results



Hierarchical PBSMT with Reordering



- Training corpus is augmented with English - Hindi bilingual dictionary
- Dictionary augmented corpus is reordered at source side to improve syntactic order
- Hierarchical Phrase-based SMT is trained on the resulting corpus from above steps
  - **Can capture the longer phrases**
  - **Can re-order long-distance phrases**

Sen, S., Banik, D., Ekbal, A., & Bhattacharyya, P.. IITP English-Hindi Machine Translation System at WAT 2016. Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 216–222, Osaka, Japan, December 11-17 2016.

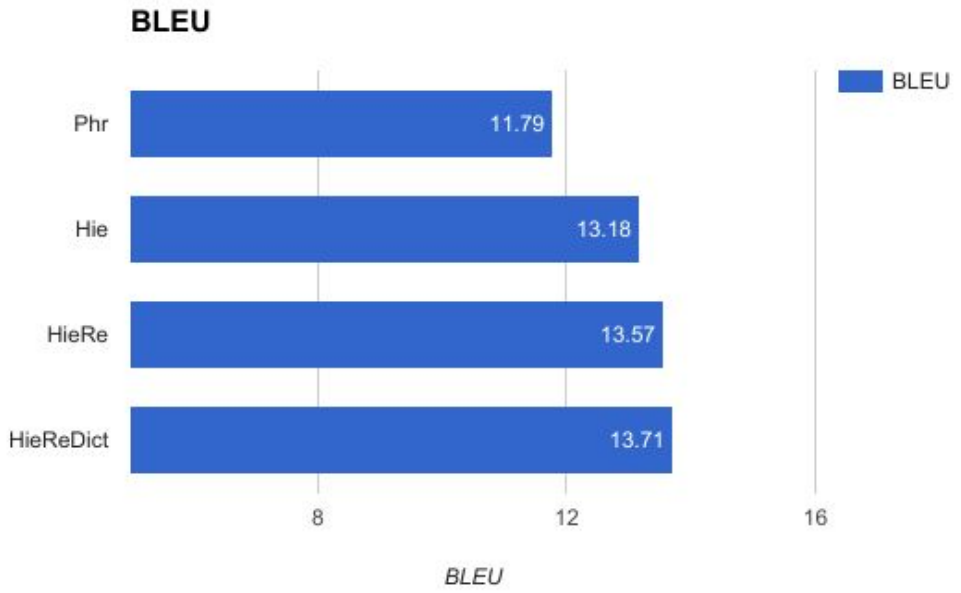
Hierarchical PBSMT with Reordering: Experimental Setup

- Training using Moses Toolkit [1]
- Stanford parser [2] for parsing source sentences
- Source-side reordering using CFILT pre-ordering tool [3]
- GIZA++ [4] for word-alignment
- 4-gram language model with modified Kneser-Ney smoothing using KenLM
- Distortion limit 6 (only for baseline Phrase-based SMT)
- Minimum-Error-Rate-Training (MERT) for tuning

1. https://github.com/moses-smt/mosesdecoder
2. http://nlp.stanford.edu/software/lex-parser.html
3. http://www.cfilt.iitb.ac.in/~moses/download/cfilt_preorder
4. https://github.com/moses-smt/giza-pp

Sen, S., Banik, D., Ekbal, A., & Bhattacharyya, P.. **IITP English-Hindi Machine Translation System at WAT 2016.** Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 216–222, Osaka, Japan, December 11-17 2016.

Hierarchical PBSMT with Reordering



Sen, S., Banik, D., Ekbal, A., & Bhattacharyya, P.. IITP English-Hindi Machine Translation System at WAT 2016. Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 216–222, Osaka, Japan, December 11-17 2016.

Output examples

Source: the rain and cold wind on Wednesday night made people feel cold.

Rerfence: बुधवार रात को हुई बारिश व ठंड हवा ने लोगों को खूब ठंड का एहसास कराया।

Output examples

Source: the rain and cold wind on Wednesday night made people feel cold.

Rerfence: बुधवार रात को हुई बारिश व ठंड हवा ने लोगों को खूब ठंड का एहसास कराया।

Hie: बुधवार रात को बरसात और ठंडी हवा **ने** लोग को ठंड लग रह थी।

(budhavAra rAta ko barasAta aura ThaMDI havA ne logoM ko ThaMDa laga rahI thI.)

HieRe: बुधवार की रात को बरसात और ठंडी हवा **से** ठंडक महसूस हुई।

Old to English to Modern English NMT

- Languages are changing with passage of time: orthography, spelling variations, writing style
- **For example**,
  - He cwæð, "Gif ge forgyfað, eow bið forgyfen.” (from 9th century)
  - He said, "If ye forgive, ye shall be forgiven." (from 18th century)
- Necessary to rewrite the old text for modern readers
- Posed this rewriting of old text into modern version as a NMT problem
- Old-modern parallel texts are not abundant (2.7K only for this task)
- Extracted parallel data from the original data

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, Andy Way (2019). *Take Help from Elder Brother: Old to Modern English NMT with Phrase Pair Feedback.* In 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019), April 2019, La Rochelle, France, In press.

Our Approach

- Assuming every source phrase *e* is aligned to a set of target phrase $F = (f_1, f_2, ..., f_n)$, we choose:
  - all parallel phrases from the phrase table of a SMT
  - phrase pairs (e, f_t) with $P(f_t|e) \geq 0.5$
  - phrase pairs (e, f_t) with $P(f_t|e) = 1$

- **Augmented Corpus** = *N* × *Original corpus* + *Extracted phrase pairs*
- *N*: ratio of the number of extracted phrase pairs to number of original parallel sentences

Dataset

- The Homilies of the Anglo-Saxon Church by Ælfric of Eynsham (c.950 – c.1010) as old English and its translation by Benjamin Thorpe (c.1782 – c.1870) as modern English

- Monolingual Bible (modern version) corpus for comparing our approach with back-translation

	Train	Dev	Test	Bible
#Sent	2716	500	500	31,102

Experimental Setup

- **Framework:** Attention-based encoder-decoder
- **Hyperparameters:**
  - Embedding size: 128
  - Hidden size: 256
  - Mini-batch size: 40
  - Maximum sentence length: 80 words
  - Adam optimizer
- For generating phrase table, we use Moses SMT tool

Results

SYSTEM		Data Size	Vocab		BLEU	METEOR	TER
			old	mod			
	PBSMT	2,716	8,878	5,102	39.95	36.96	37.99
	Baseline-NMT (B_N)	2,716			**10.03**	**15.95**	**90.06**
Back-Trans	B_N+10k *BT*	12,716	15,067	10,948	21.90	24.95	64.66
	B_N+20k *BT*	22,716	17,341	13,162	24.23	25.83	58.15
	B_N+*BT*	33,818	19,083	14,859	**29.10**	**30.70**	**52.48**
Proposed Approach							
Type-A	B_N+$PHR_{Org,p=1.0}$	341,659	8,878	5,102	20.83	25.84	84.66
	B_N+$PHR_{Org,p\geq 0.5}$	385,015			25.41	27.79	69.42
	B_N+PHR_{Org}	485,739			**28.76**	**28.56**	**56.37**
Type-B	B_N+$PHR_{Org+BT,p=1.0}$	4,850,270	19,080	14,855	25.30	26.50	63.33
	B_N+$PHR_{Org+BT,p\geq 0.5}$	5,094,325			27.93	28.76	60.58
	B_N+$PHR_{Org.+BT}$	6,068,402			25.17	26.95	68.76
Type-C	B_N+*BT*+PHR_{Org}	512,613	19,083	14,859	**32.35**	**31.88**	**50.36**
	B_N+*BT*+PHR_{Org+BT}	6,073,265			**29.37**	**30.31**	**56.02**

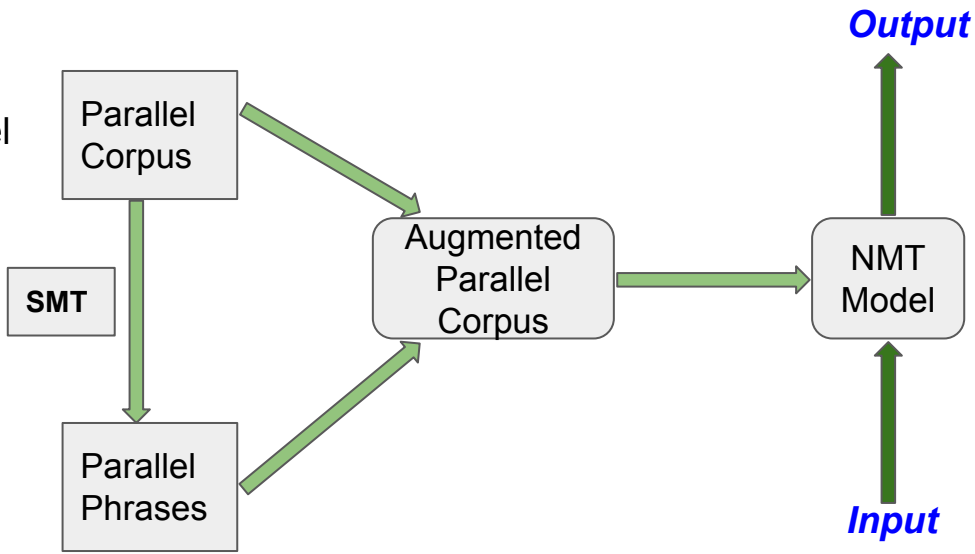
BT: Back-translated Bible Copus;

Org: Original Corpus;

$PHR_{Org,p\geq 0.5}$: Phrase pairs (having probability ≥0.5) from original corpus.

Neural Machine Translation of Low-resource Languages using SMT Phrase Pair Injection

- An SMT-NMT hybrid system
- Augmented original parallel sentences with parallel phrases from the phrase table
- Used augmented parallel corpus to train the NMT model
- Parallel phrases enrich the training data to help low-resource NMT model



*Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. Neural machine translation of low-resource languages using SMT phrase pair injection. Natural Language Engineering, Cambridge University Press, pages 1–22, June 2020

Low resource NMT with SMT (Cont...)

- True low-resource language pairs: Hindi-English, Bengali-Hindi
  - Heath, Tourism, Judicial domain
- Attention based GRU, Transformer
- Data size: 5-23 K
- Translation for 10 more directions/language pairs
- Low resource NMT > SMT

Neural Machine Translation of Low-resource Languages using SMT Phrase Pair Injection

- Using available parallel sentences, train source-target SMT
- From generated phrase table, extract phrase pairs having translation probability over certain threshold (>= 0.5)
- Augment the original parallel corpus with the extracted source-target phrase pairs
- Train **source ⇒ target NMT** model using the augmented corpus

Data Statistics and Results

- **Training**

Language Pair	Hindi-English			Hindi-Bengali	
Domain	Health	Tourism	Judicial	Health	Tourism
Size	25,000	25,000	5561	25,000	25,000

- **BLEU Score**
  - **H:** Health, **T:** Tourism, **J:** Judicial

	Hindi→English			English→Hindi			Hindi→Bengali		Bengali→Hindi	
	H	T	J	H	T	J	H	T	H	T
Baseline	22.08	20.41	25.34	20.71	17.44	21.84	22.81	18.90	24.92	22.92
Proposed	**24.65**	**24.40**	**29.43**	**23.97**	**19.74**	**25.50**	**26.17**	**23.47**	**28.90**	**26.88**

Results

Model	Hindi→English			English→Hindi			Hindi→Bengali		Bengali→Hindi		Old→Mod
	H	T	J	H	T	J	H	T	H	T	English
SMT	23.07	24.39	29.36	20.64	19.24	26.75	28.50	25.09	29.81	29.62	39.95
Attention-based Encoder-Decoder											
Baseline	14.02	12.14	6.97	14.04	12.26	11.25	18.11	13.81	21.2	19.04	10.03
+ $Set_{p\geq 0.5}$	18.53	**20.79**	18.21	14.88	15.95	**20.69**	20.39	**19.25**	21.64	20.46	25.41
+ $Set_{p=1.0}$	**19.43**	19.75	**22.33**	**16.35**	**16.67**	17.19	20.70	18.76	**22.58**	**22.06**	20.83
+ Set_{all}	18.73	18.22	21.39	16.14	14.92	19.74	**20.93**	18.08	21.45	21.42	**28.76**
▲	5.41 ↑	8.65 ↑	15.36 ↑	2.31 ↑	4.41 ↑	9.44 ↑	2.82 ↑	5.44 ↑	1.38 ↑	3.02 ↑	18.73 ↑
Transformer											
Baseline	22.08	20.41	25.34	20.71	17.44	21.84	22.81	18.90	24.92	22.92	27.94
+ $Set_{p\geq 0.5}$	**25.70***	23.59	**29.85***	22.24	19.66	**25.99**	**27.04**	**24.00**	28.74	26.42	33.40
+ $Set_{p=1.0}$	24.65	**24.40***	29.43	**23.97***	**19.74***	25.50	26.17	23.47	**28.90**	**26.88**	**32.67**
+ Set_{all}	25.63	23.80	29.19	23.01	18.19	25.53	26.03	23.73	28.51	26.86	33.61
▲	3.68 ↑	3.99 ↑	4.51 ↑	3.26 ↑	2.30 ↑	4.15 ↑	4.23 ↑	5.10 ↑	3.98 ↑	3.96 ↑	5.67 ↑

Set$_{all}$: set of all parallel phrases from the phrase table; ***Set$_{p\geq 0.5}$***: set of phrase pairs (e, ft) with $p(f_t|e) \geq 0.5$; ***Set$_{p=1}$***: set of phrase pairs (e, ft) with $p(f_t|e) = 1$; **H**: Health, **T**: Tourism, **J**: Judicial; *****: Better than SMT (though in general, SMT performs better than NMT in low resource settings)

Effect on Subword

System	BLEU Score
Subword (Sennrich et al.)	21.58
Our Approach with $Set_{p=0.5}$	23.27
Our Approach with $Set_{p=1.0}$	22.33
Our Approach with Set_{all}	23.18

- *Our approach using Transformer at sub-word level for English-Hindi direction on Health domain*

Comparative Systems

System	BLEU Score
SMT	20.64
Transformer	20.71
Pre-translation (Niehues et al.)	20.40
Mixed Pre-translation (Niehues et al.)	19.58
Transformer with Back-translation	18.75
Proposed best model	23.97

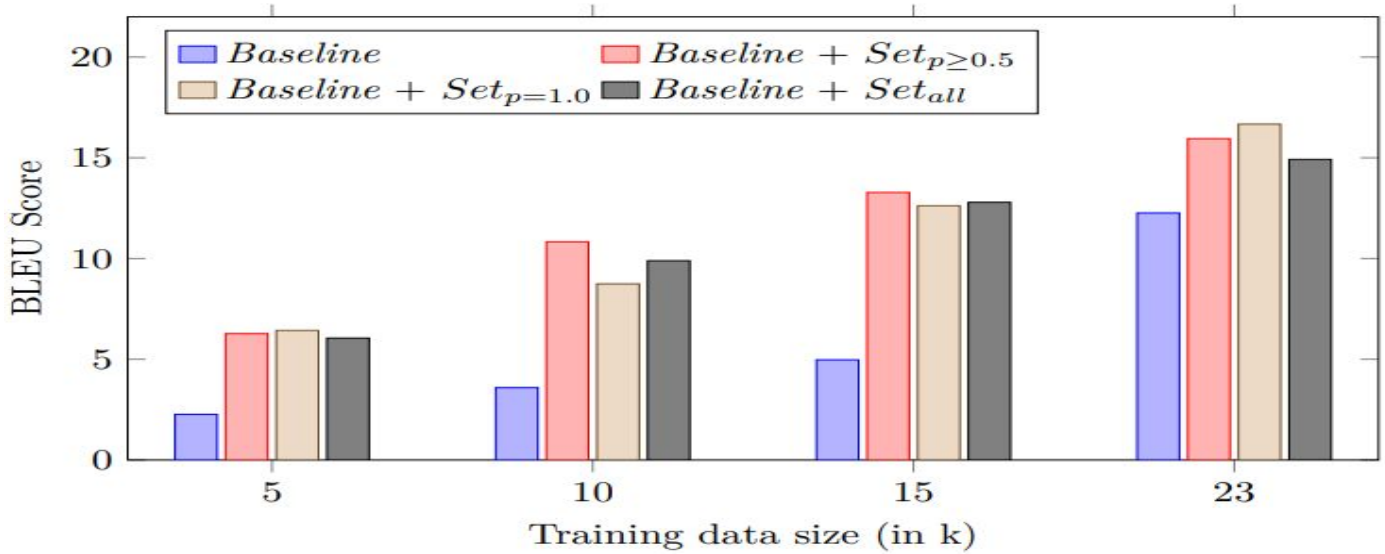
- *Comparative systems for English-Hindi for Health domain*

Comparative Systems

System	BLEU Score
PhraseNet (embedding = 620 and hidden = 1000)	9.62
PhraseNet (embedding = 128 and hidden = 256)	12.05
PhraseNet (embedding = 300 and hidden = 600)	12.65
Our Approach with $Set_{p>=0.5}$	15.95
Our Approach with $Set_{p=1.0}$	16.67
Our Approach with Set_{all}	14.92

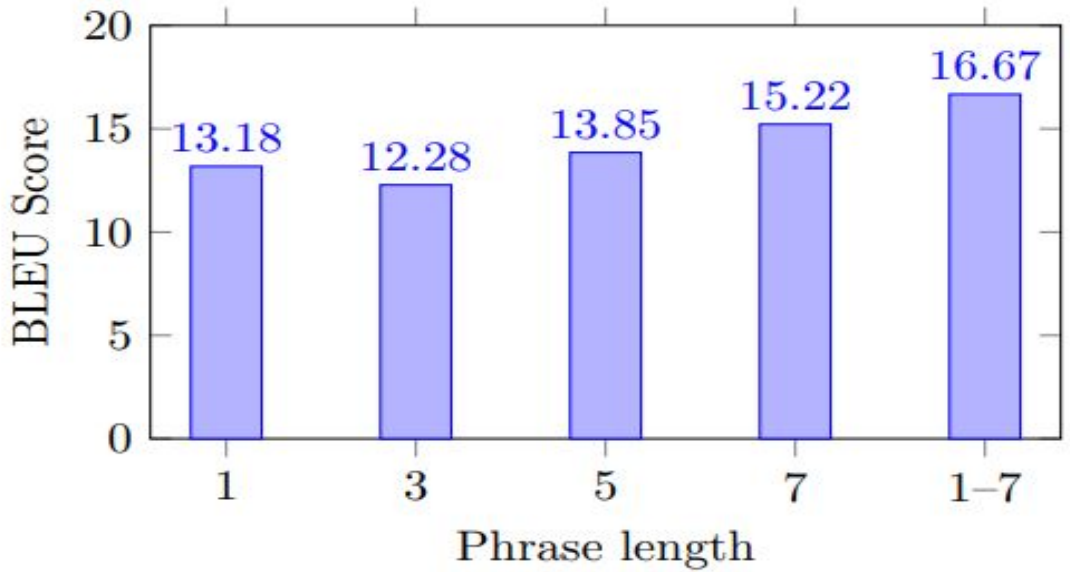
- *Comparison of our approach (using attention-based GRU) with PhraseNet for English-Hindi for Health domain. In parenthesis, we show the dimension*

Effect of Training Data Size



Comparison of different NMT models with incremental original training data

Effect of Phrase Length



Performance of proposed method with different phrase lengths

Multilingual NMT

- MT models are large on disk and have huge memory footprint and requires huge computing power (CPU, GPU)
  - Considerably higher deployment cost and/or management
  - Multilingual training would be problematic
- Multilinguality is less explored with SMT
  - SMT was difficult to train for the multilingual models
- NMT is well suited for multilingual models
- ***Most importantly*** -- sharing parameters across multiple languages helps low-resource languages

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. IITP-MT at WAT 2018: Transformer-based Multilingual Indic-English Neural Machine Translation System. Proceedings of the 5th Workshop on Asian Translation (WAT, co-located with PACLIC 2018), Hong Kong, China, December. 2018.

Multilingual Supervised NMT

- Multilingual Neural Machine translation between English and Indic languages
- Transformer-based (Vaswani et al. 2017) ***many-to-one*** and ***one-to-many*** models
- Our systems ranked top in three out of four human evaluation scheme

Our Approach

- Pre-process using Byte-pair Encoding (BPE) based subword
- **Multilingual models**: additional token indicating which Indic language a sentence pair belongs to is added at the beginning of every source sentence. e.g.,
  - HI## I need your help. मुझे आपकी मदद चाहिए ।
  - BN## Where the mind is without fear! চিত্ত যেথা ভয় শূন্য!

- Combine all training data to train a single model to translate between all the languages

Dataset

Language Pair	#Sentences
Bengali (BN) - English	337,428
Hindi (HI) - English	84,557
Malayalam (ML) - English	359,423
Tamil (TA) - English	26,217
Telugu (TE) - English	22,165
Urdu (UR) - English	26,619
Sinhalese (SI) - English	521,726

Number of Subword Merge

Data Pair	Source			Target		
	Original Vocab	Merge	Final Vocab	Original Vocab	Merge	Final Vocab
BN-EN	90,482	8,000	8,394	56,498	5,000	5,248
HI-EN	24,470	4,000	4,286	24,380	4,000	4,150
ML-EN	253,360	10,000	10,351	58,320	5,000	5,273
SI-EN	169,603	9,000	9,392	72,093	7,000	7,417
TA-EN	18,723	3,500	3,675	18,723	2,000	2,114
TE-EN	12,728	2,000	2,230	9,929	1,500	1,633
UR-EN	13,581	3,000	3,268	12,854	2,000	2,126

- *Original vocabulary size, number of BPE merge and final vocabulary size after applying BPE for each training data pair. We decided the BPE merge values without any rigorous exploration.*

Experimental Setup

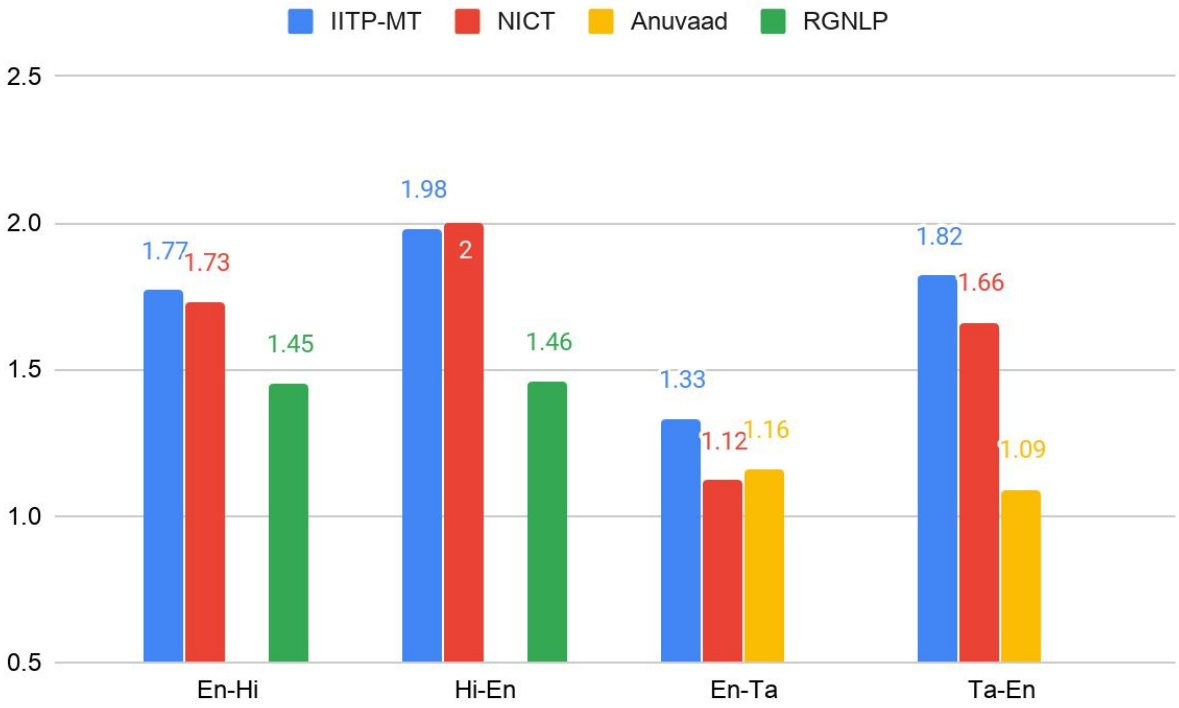
- Transformer network (Vaswani et al. 2017)
- Embedding dimension = 512
- Hidden dimension = 512
- Learning rate = 0.0002
- Dropout rate = 0.2
- Adam optimizer
- Minibatch size of 2000 words
- Maximum sentence length 50 tokens

Results

System	Bilingual model	Multilingual model	Improvement
BN → EN	18.24	20.05	+1.81
HI → EN	27.11	32.95	+5.84
ML → EN	10.56	19.94	+9.38
SI → EN	18.22	21.35	+3.13
TA → EN	11.58	22.42	+10.84
TE → EN	16.15	30.96	+14.81
UR → EN	20.02	26.56	+6.54
EN → BN	13.38	13.27	-0.11
EN → HI	24.25	26.60	+2.35
EN → ML	20.92	13.50	-7.42
EN → SI	12.75	10.64	-2.11
EN → TA	11.88	18.81	+6.93
EN → TE	14.21	25.81	+11.60
EN → UR	18.73	21.48	+2.75

- ***For Indic-to-English***: multilingual model performs better than the separate bilingual models for all
- ***For English-to-Indic***: multilingual model performs better than bilingual models only for relatively low-resource languages

Human Evaluation



References

- Sen, Sukanta, K. Gupta, A. Ekbal and P. Bhattacharyya (2019). Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL* 2019), pp. 3083-3089.
- Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. Neural machine translation of low-resource languages using SMT phrase pair injection. Natural Language Engineering, Cambridge University Press, pages 1–22, June 2020
- Sen, S., et al. (2019). Take help from elder brother: Old to modern english nmt with phrase pair feedback. *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*. 2019.
- Sen, Sukanta, Asif Ekbal, and Pushpak Bhattacharyya (2019). Parallel corpus filtering based on fuzzy string matching. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. 2019.
- Gupta, K. K., Haque, R., Ekbal, A., Bhattacharyya, P., & Way, A. (2020). Syntax-informed interactive neural machine translation. In *Proceedings of IJCNN 2020*.
- Niehues, Jan, et al. Pre-Translation for Neural Machine Translation (2016). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.

References

- Banerjee T, Kunchukuttan A, Bhattacharya P. Multilingual Indian Language Translation System at WAT 2018: Many-to-one Phrase-based SMT. In WAT@ PACLIC 2018
- Rudra Murthy, Anoop Kunchukuttan, Pushpak Bhattacharyya. Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages. In NAACL 2019.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.
- Vaswani, Ashish, et al. "Attention is all you need. *Advances in neural information processing systems*. 2017.
- Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics* 19.2 (1993): 263-311.
- Kim, Yunsu, et al. "Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.
- Zoph, Barret, et al. "Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016.
- Artetxe, Mikel, et al. "Unsupervised neural machine translation. In proceedings of the *6th International Conference on Learning Representations, ICLR 2018*. 2018.

References

- Lample, Guillaume, et al. "Phrase-Based & Neural Unsupervised Machine Translation." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.
- Britz, Denny, Quoc Le, and Reid Pryzant. "Effective domain mixing for neural machine translation." *Proceedings of the Second Conference on Machine Translation*. 2017.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016.
- Johnson, Melvin, et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation." *Transactions of the Association for Computational Linguistics* 5 (2017): 339-351.
- Desai, Pratik, Amit Sangodkar, and Om P. Damani. "A domain-restricted, rule based, english-hindi machine translation system based on dependency parsing." *Proceedings of the 11th International Conference on Natural Language Processing*. 2014.
- Naskar, Sudip Kumar, and Sivaji Bandyopadhyay. "A phrasal EBMT system for translating English to Bengali." *Satellite Workshop*. 2006.
- Ambati, Vamshi, and U. Rohini. "A hybrid approach to example based machine translation for Indian languages." *ICON* 2007 (2007): 5th.
- Koehn, Philipp, and Hieu Hoang. "Factored translation models." *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 2007.
- Ramanathan, Ananthakrishnan, et al. "Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009.

References

- Singh, Thoudam Doren, and Sivaji Bandyopadhyay. "Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations." *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*. 2010.
- Sen, Sukanta, et al. "IITP-MT at WAT2018: Transformer-based Multilingual Indic-English Neural Machine Translation System." *WAT@ PACLIC*. 2018.
- Murthy, Rudra, Anoop Kunchukuttan, and Pushpak Bhattacharyya. "Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.
- K. Gupta, R. Haque, A. Ekbal, P. Bhattacharyya, A. Way (2020). Syntax-informed interactive neural machine translation. IJCNN 2020
- Banik, Debajyoty, et al. "Can smt and rbmt improve each other's performance?-an experiment with English-Hindi translation." *Proceedings of the 13th international conference on natural language processing*. 2016.
- Sen, Sukanta, et al. "IITP English-Hindi machine translation system at wat 2016." *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*. 2016.

References

- Sinha, R., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., and Jain, A. (1995). *ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages*. In IEEE International Conference on Systems, Man and Cybernetics
- R.M.K. Sinha, ‘R & D on Machine Aided Translation at IIT Kanpur: ANGLABHARTI and ANUBHARTI Approaches’, Invited paper at Convention of Computer Society of India, (CSI’96), Bangalore, 1996
- Bharati A, Chaitanya V, Kulkarni AP, Sangal R, Rao GU. ANUSAARAKA: overcoming the language barrier in India. arXiv preprint cs/0308018. 2003 Aug 7.
- Ramanathan, A., Hegde, J., Shah, R., Bhattacharyya, P., and Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In IJCNLP
- Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation
- Anoop Kunchukuttan, Pushpak Bhattacharyya. Orthographic Syllable as basic unit for SMT between Related Languages. In EMNP 2016
- Dave, S., Parikh, J., & Bhattacharyya, P. (2001). Interlingua-based English–Hindi machine translation and language divergence. *Machine Translation*, *16*(4), 251-304.
- Jha GN. The TDIL Program and the Indian Language Corpora Intitiative (ILCI). In LREC 2010 May 17.

Some Resource Links

For Indian Language MT systems: https://www.aclweb.org/anthology/O13-2003.pdf

Philipp Koehn MT class page from JHU: http://mt-class.org/jhu/syllabus.html

NLP for ILs:
https://anoopkunchukuttan.gitlab.io/publications/presentations/wildre_keynote_2020.pdf

SMT between related languages:
https://anoopkunchukuttan.gitlab.io/publications/presentations/naacl-2016-tutorial.pdf

Thank You for Your Attention!