CS626: Speech, NLP and Web

Machine Translation- Intro and Paradigms

Pushpak Bhattacharyya and Sourabh Deoghare

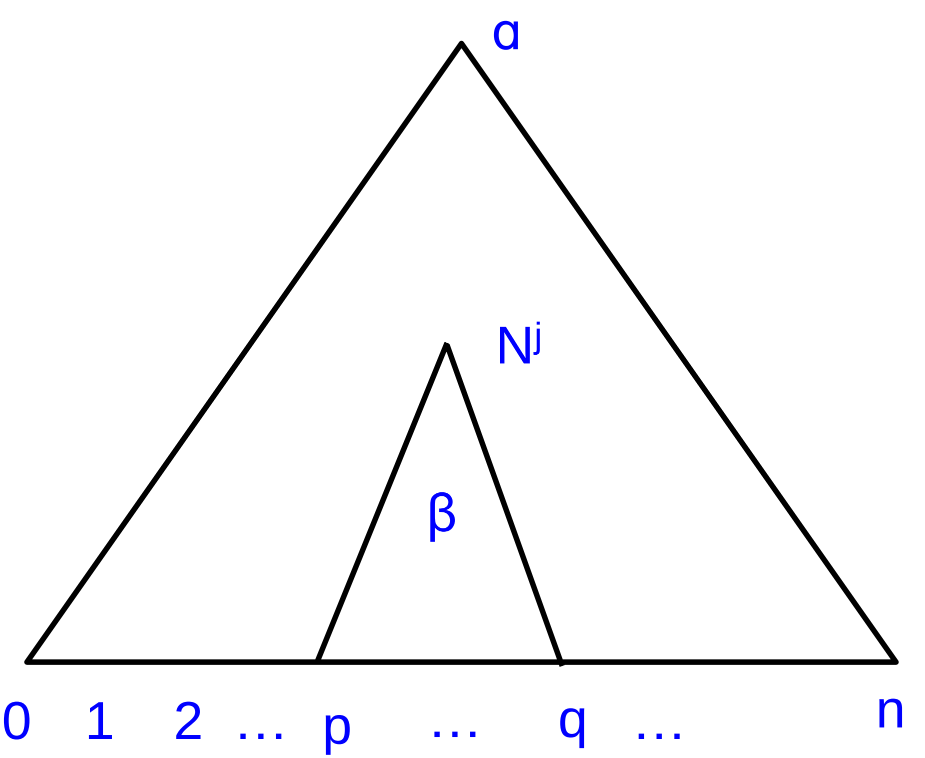
Computer Science and Engineering Department

IIT Bombay

Week 8 of 23rd September, 2024

1-slide recap of week of 9th Sep

- Domination
- Probabilistic Parsing:
  - *T*= argmax [P(T|S)]*
- Probability of a sentence
  - = *P(w0,l)=Σt P(t)*



$$\beta_j(p,q) = P(W_{p-q} \mid N^j_{pq})$$
$$= \sum_{k,r,l} P(N^j \rightarrow N^k\ N^l).P(W_{p-r} \mid N^k_{pr}).P(W_{r-q} \mid N^l_{rq})$$
$$= \sum_{k,r,l} P(N^j \rightarrow N^k\ N^l).\beta_k(p,r).\beta_l(r,q)$$

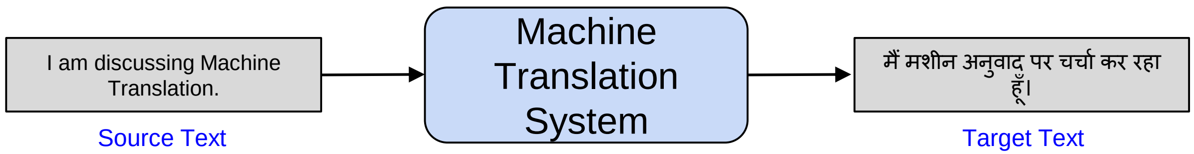
$$\delta_i(p,q) = \max{}_{j,r,k}\ P(N^i \rightarrow N^j\ N^k).\delta_j(p,r).\delta_k(r,q)$$

Stress Test for Parsing:
A very difficult parsing situation!-
the buffalo sentence

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \prod_{k=2}^{n} \frac{k}{n+k}, \quad n \geq 0$$

Introduction to Machine Translation (MT) (1/2)

- What is Machine Translation?
  - Translation of a piece of text in one language into another through a computer program.
  - The target text should convey the ***exact*** meaning as the source text.



- Why do we need Machine Translation?
  - To reduce/remove the language barrier.

- Who needs it?
  - Communication, Travel, Entertainment, Administration, Education, Industry, etc.

Introduction to Machine Translation (MT) (2/2)

- Why are we interested?
  - Any multilingual NLP system involves MT as some level
  - Challenging and old problem
  - Deals with Natural Language Understanding and Generation
  - MT theories and techniques has applicability in a range of other NLP problems

- Why is MT a difficult problem? **Ambiguity**
  - Different languages have different properties
  - Different word order
  - Polysemy and synonymy
  - Morphological Richness

What does MT need to do?

Deal with ***Language Divergence***

Language Divergence

- Languages express meaning in divergent ways.

- Syntactic Divergence:
  - Arises because of the ***difference in structure***

- Lexical-Semantic Divergence:
  - Arises because of difference in ***semantic properties*** of languages

Types of Syntactic Divergence

- Constituent Order Divergence:

English: He is waiting for him.
Hindi: वह उसके लिए इंतजार कर रहा है।

Subject	He	वह
Verb	waiting	इंतजार कर रहा है
Object	him	उसके

- Adjunction Divergence:

English: Delhi, the capital of India, has many historical buildings.
Hindi: भारत की राजधानी दिल्ली में बहुत सी एतिहासिक इमारतें हैं

- Null Subject Divergence:

English: I am going.
Hindi: जा रहा हूँ।।

Types of Lexical-Semantic Divergence

- Conflational Divergence:

English: He stabbed him.
Hindi: उसने उसे छुरे से मारा

- Categorial Divergence (Lexical Category Change):

English: They are competing.
Hindi: वे प्रतिस्पर्धा कर रहे हैं

- Head-swapping Divergence (Promotion or Demotion of a Logical Modifier):

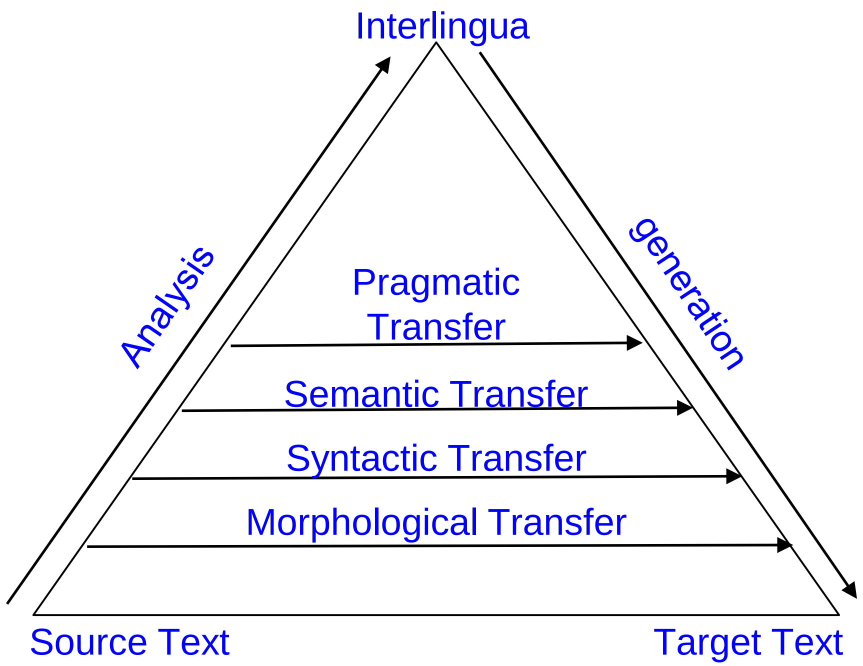
English: The play is on.
Hindi: खेल चल रहा है

Conceptual Model of MT: Vauquois Triangle

- Problems:
  - Word-level Transfer:

I am eating an apple.

मैं हूँ खा रहा एक सेब।

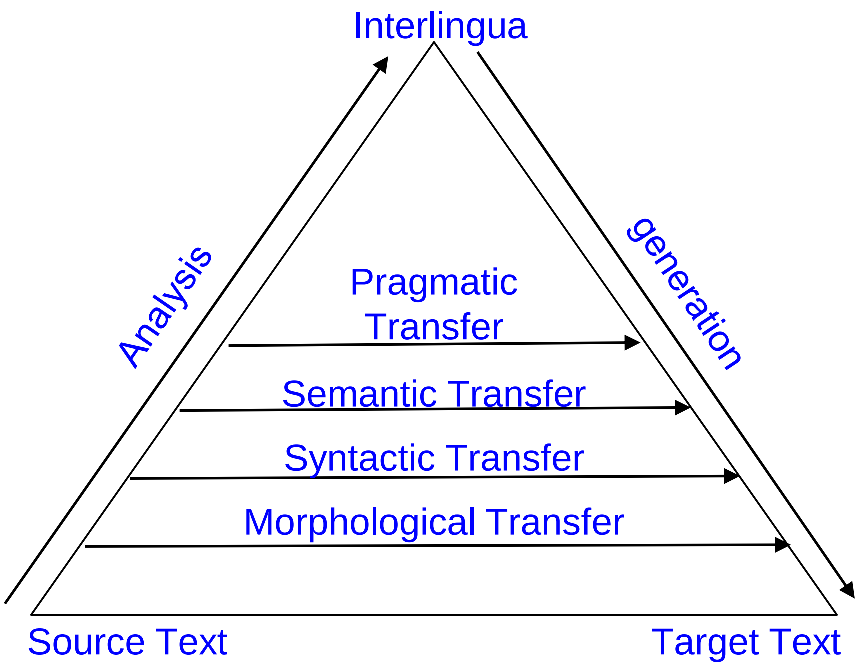


Conceptual Model of MT: Vauquois Triangle

- Problems:
  - Syntax-level Transfer:

I miss the bus every day.

मैं हर दिन बस को याद करता हूँ।

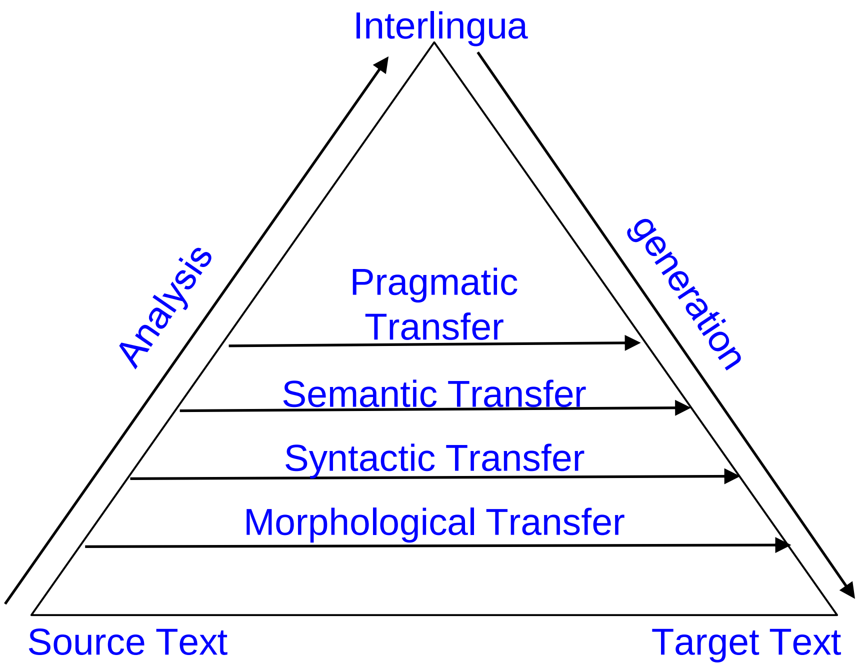


Conceptual Model of MT: Vauquois Triangle

- Problems:
  - Semantic-level Transfer:

His departure was for good.

उसका प्रस्थान अच्छे के लिए था।



Approaches

- Rule-based and Knowledge-based MT:
  - Interlingua-based
  - Transfer-based

- Limitations:
  - Costly in terms of time and money
  - Highly complex
  - Adaptation is difficult

- Data Driven MT: Parallel Corpus - contains tuples of (source, translation)
  - Example-based MT
  - Statistical MT
  - Neural MT

Paradigms of Machine Translation

Human effort

Data hungry

RBMT → EBMT → SMT → NMT

(Rule based machine translation) (Example based machine translation) (Statistical machine translation) (Neural machine translation)

Rules handmade by human expert → Rules by learning from data

Topics To Be Discussed

- Machine Translation (MT)
  - Introduction
  - **Statistical Machine Translation**
  - Neural Machine Translation
  - Evaluation

Problem Formulation

- Goal: Given a foreign sentence *f*, find the most likely English translation *e.*
- Probabilistic Model:
  - We would like to have a measure of confidence for the translations we learn.
  - We would like to model uncertainty in translation.

- Notations:
  - Source language (*F*)
  - Target language (*E*)
  - Target language sentence (*f*)
  - Source language sentence (*e*)

$$\bar{e} = \arg\max_{e} P(e|f)$$

Noisy Channel Model (1/2)

- Sees translation as a process of recovering the original sentence given the corrupted sentence.

- Decomposes $P(e|f)$ into $P(f|e) * P(e)$ / $P(f)$

- Steps:
  - Generate *e* using $P(e)$
  - Pass *e* through the channel
  - Get *f*, a corruption of *e*

- Why do we do it?
  - Makes it easier to mathematically represent translation and learn probabilities
  - Allows to separately model **Adequacy** and **Fluency**

Noisy Channel Model (2/2)

- Language Model:
  - How likely is *e* to be an English sentence?: $P(e)$
  - Monolingual data

- Translation Model:
  - How likely is *f* to be a translation of *e*: $P(f|e)$
  - Bilingual data

- Generative Modelling:
  - Generate *e* with probability $P(e)$
  - Pass *e* through noisy channel and get *f* with probability $P(f|e)$
  - Given *f*, what is the best translation e: $\mathrm{argmax}\ P(e|f)$

Language Model

- Given an English sentence e with words in order $e_1, e_2, \ldots, e_l$:

  $$P(e) = P(e_1, e_2, \ldots, e_l)$$
  $$= P(e_1) * P(e_2 | e_1) * \ldots * P(e_l | e_1, e_2, \ldots, e_{l-1})$$

- N-Gram Language Model:
  - Let's assume that $P(e_i)$ depends on previous N-1 words only.
  - $P(e_i | e_1, e_2, \ldots, e_{l-1}) = P(e_i | e_{i-N}, e_{i-N+1}, \ldots, e_{i-1})$

- If N = 2, it is a Bigram model

  P(I am discussing Statistical Machine Translation)

  *= P(I | START) * P(am | I) * …. * P(END | Translation)*

Translation Model

- How do we learn *P(f|e)?*

- We first need to learn word-level translation probabilities to learn the sentence-level translation probabilities.

- English sentence $e = e_1, e_2, \ldots, e_l$
- Foreign sentence $f = f_1, f_2, \ldots, f_m$

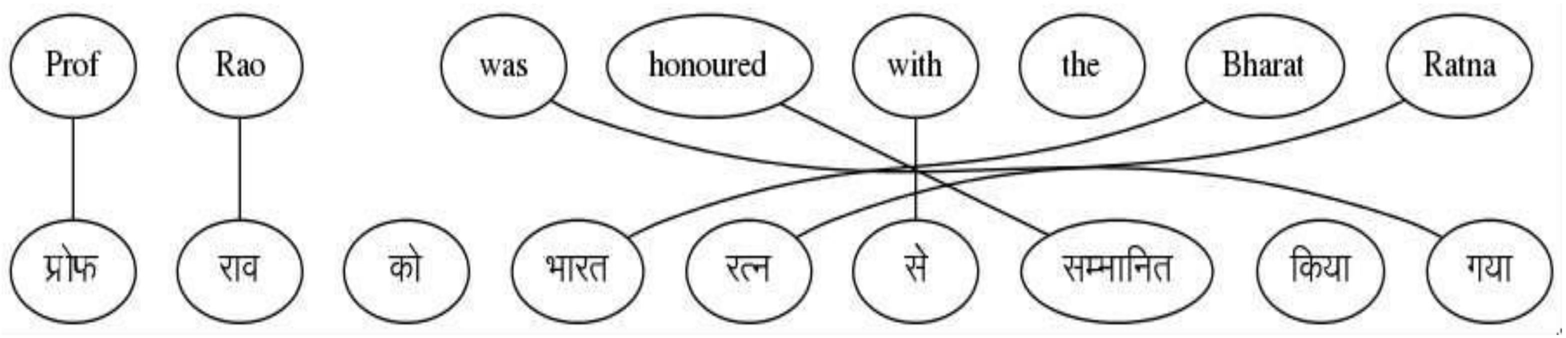
Translation Model: Co-occurrence

- Words which occur together in parallel sentence pairs are likely to be translations (higher *P(f|e)*)

Parallel Corpus	
A boy is **sitting** in the kitchen	एक लडका रसोई में **बैठा** है
A boy is playing **tennis**	एक लडका **टेनिस** खेल रहा है
A boy is **sitting** on a round table	एक लडका एक गोल मेज पर **बैठा** है
Some men **are watching tennis**	कुछ आदमी **टेनिस देख रहे है**
A girl is holding a black book	एक लडकी ने एक काली किताब पकडी है
Two men **are watching** a movie	दो आदमी चलचित्र **देख रहे है**
A woman is reading a book	एक औरत एक किताब पढ़ रही है
A man is **sitting** in a red car	एक आदमी एक काले कार मैं **बैठा** है

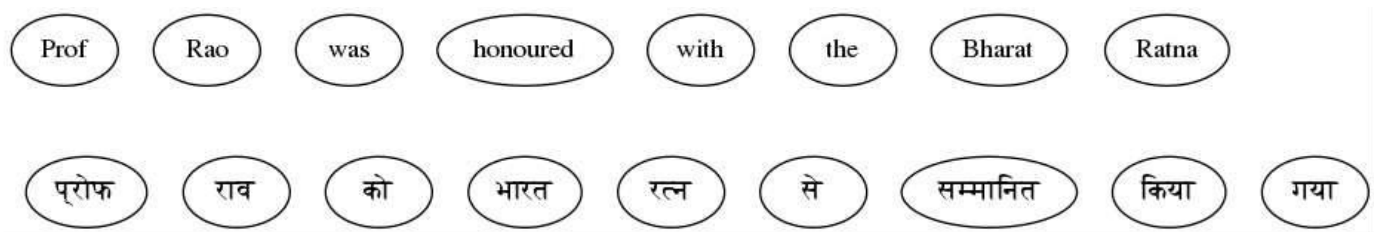
Translation Model: Alignment (1/6)

- A word in the source sentence can be aligned to a small number of words in the translation.

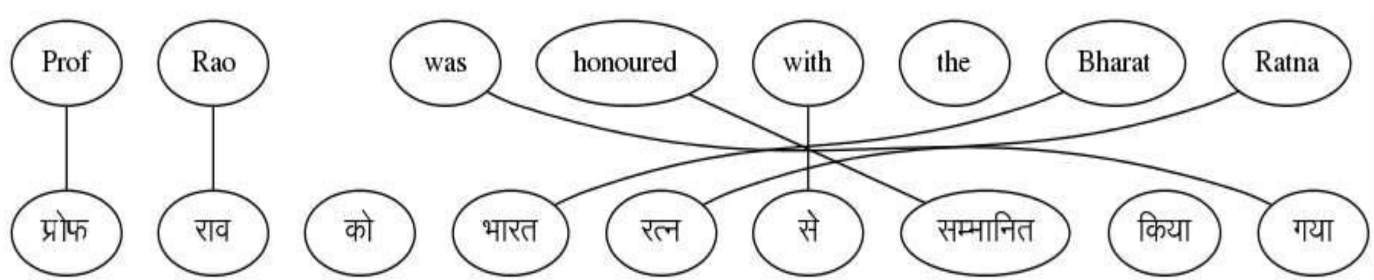


Translation Model: Alignment (2/6)

- Problem: Given a source sentence and its translation, find the word-level correspondences.



- Alignments:



Translation Model: Alignments (3/6)

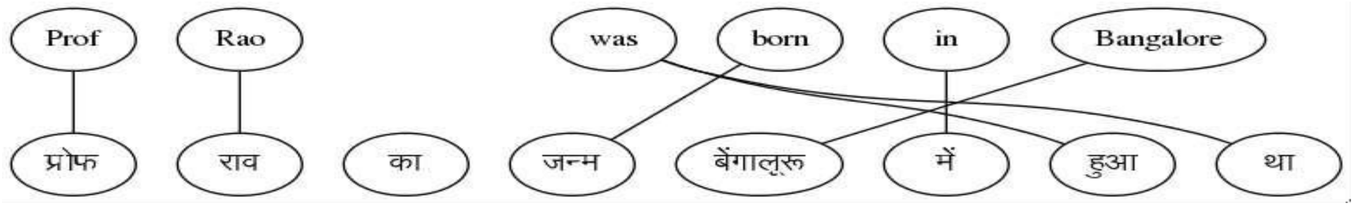
- How do we learn *P(f|e)?*

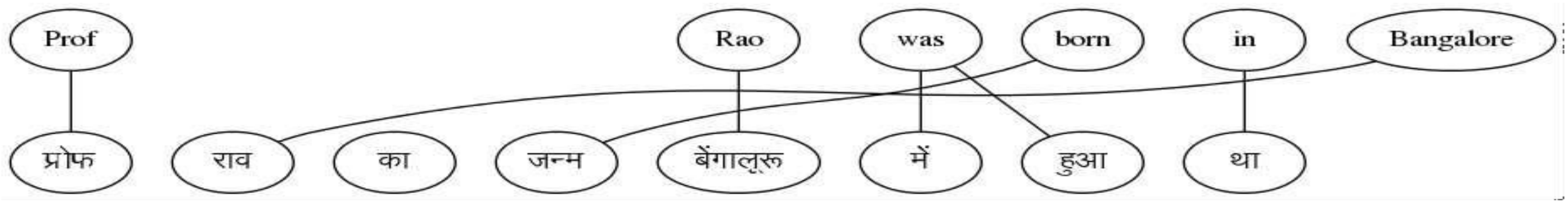
- We first need to learn word-level translation probabilities to learn the sentence-level translation probabilities.

- English sentence $e = e_1, e_2, \ldots, e_l$
- Foreign sentence $f = f_1, f_2, \ldots, f_m$
- Alignment $A = \{a_1, a_2, \ldots, a_m\}$, where a_j belongs to $\{0, 1, \ldots, l\}$.

Translation Model: Alignment (4/6)

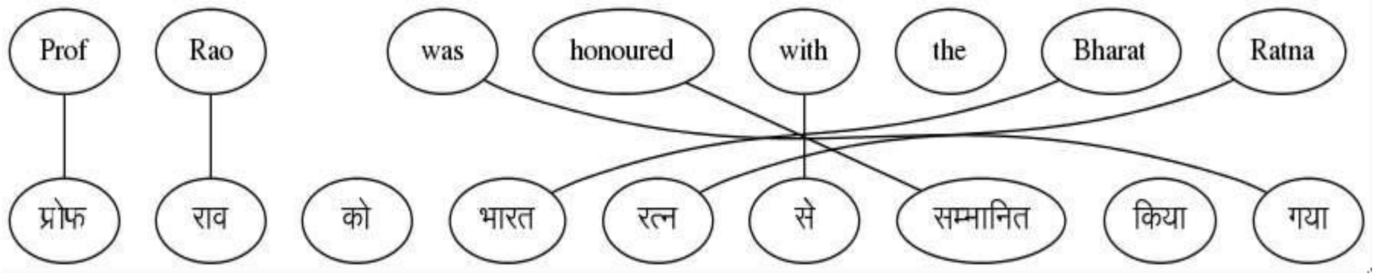
- Issue: There will be multiple possible alignments
- How many possible alignments between *e* of length *l* and *f* of length *m?*
- Can we find the correct alignments if we have multiple sentence pairs?

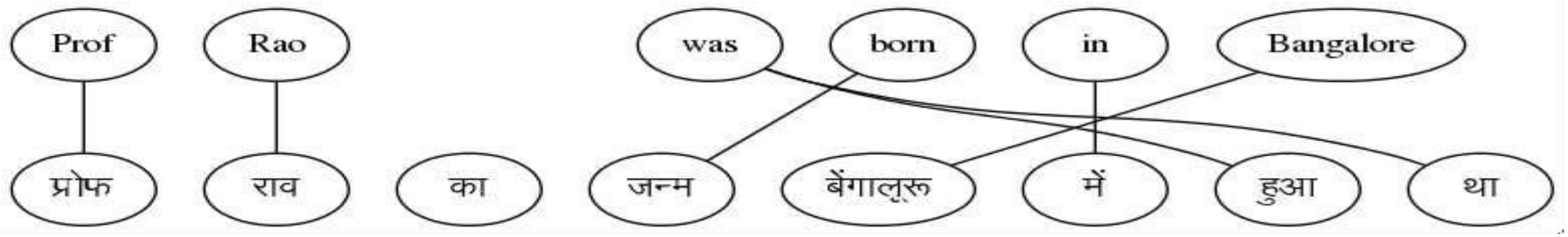




Translation Model: Alignments (5/6)

- If we knew the alignments, we could compute *P(f|e)*





- *P(f|e) = #(f, e) / #(*, e)*
  - Where *#(f, e) denotes number of times word f is **aligned** with word e*
- *P(Prof | प्रोफ) = 2/2*

Translation Model: Alignments (6/6)

- Issue: We can find the best alignment only if we know the word-level translation probabilities.
- The best alignment *a** is the one that maximizes the sentence translation probability.

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = P(a) \prod_{i=1}^{i=m} P(f_i | e_{a_i})$$

$$\downarrow$$

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}}{\operatorname{argmax}} \prod_{i=1}^{i=m} P(f_i | e_{a_i})$$

Translation Model: Expectation Maximization (EM) (1/3)

- Randomly initialize word-level translation probabilities

- Two-step iterative Process:
  - Step 1: Estimate alignment probabilities using word translation probabilities
  - Step 2: Re-estimate word translation probabilities

- As we don't know the best alignment, we consider all alignments while estimating word translation probabilities.
- Instead of taking only the best alignment, we consider all alignments and weigh the word alignments with the alignment probabilities

$$P(f|e) = \frac{expected\ \#(f,e)}{expected\ \#(*,e)}$$

Translation Model: Expectation Maximization (EM) (2/2)

- Probabilities:

  P(the | एक) = 0.7   P(house | एक) = 0.05

  P(the | घर) = 0.1   P(house | घर) = 0.8

- Alignments:

एक — the घर — house	एक — the घर house	एक the घर — house	एक the घर house
P(e, a | f) = 0.56	P(e, a | f) = 0.035	P(e, a | f) = 0.08	P(e, a | f) = 0.005
P(e, a | f) = 0.824	P(e, a | f) = 0.052	P(e, a | f) = 0.118	P(e, a | f) = 0.007

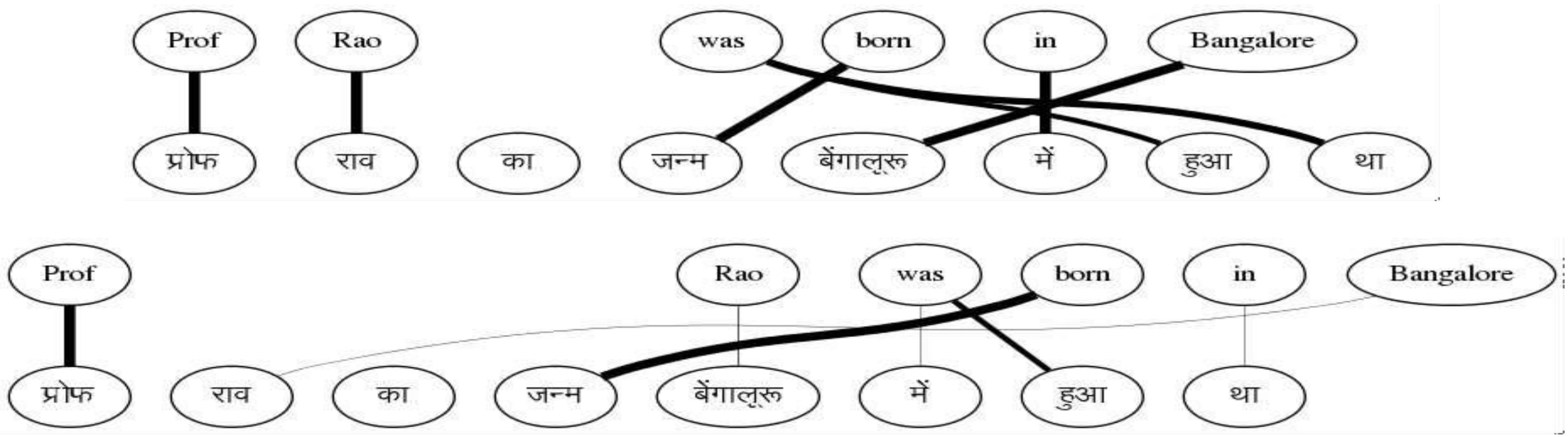
- Counts:

  P(the | एक) = 0.824 + 0.052   P(house | एक) = 0.052 + 0.007

  P(the | घर) = 0.118 + 0.007   P(house | घर) = 0.824 + 0.118

Translation Model: Expectation Maximization (EM) (3/3)

- At the end of the process:



- Note: Poor initialization may lead to convergence to local minima.

IBM Models

- IBM came up with a series of increasingly complex models
- Called IBM Models 1 to 5
- Differed in assumptions about alignment probability distributions
- Simpler models are used to initialize the more complex models
- This pipelined training helped ensure better solutions

Phrase-based SMT: Introduction

- Basic translation unit: Phrase (sequence of words), Not word
  - Note: Not necessarily linguistic phrases

- Advantages:
  - Local reordering (Intra-phrase reordering can be memorized)
    - Eg. The prime minister of India –> भारत के प्रधान मंत्री
  - Sense disambiguation based on local context (Neighbouring words help make the choice)
    - Eg. heads towards Pune → पुणे की और जा रहे है, heads the committee –> समिति की अध्यक्षता करते है
  - Institutionalized expressions, idioms can be learnt as a single unit
    - Eg. hung assembly –> त्रिशंकु विधानसभा
  - Improved fluency as a phrase could as long as an entire sentence.

Phrase-based SMT: Mathematical Model

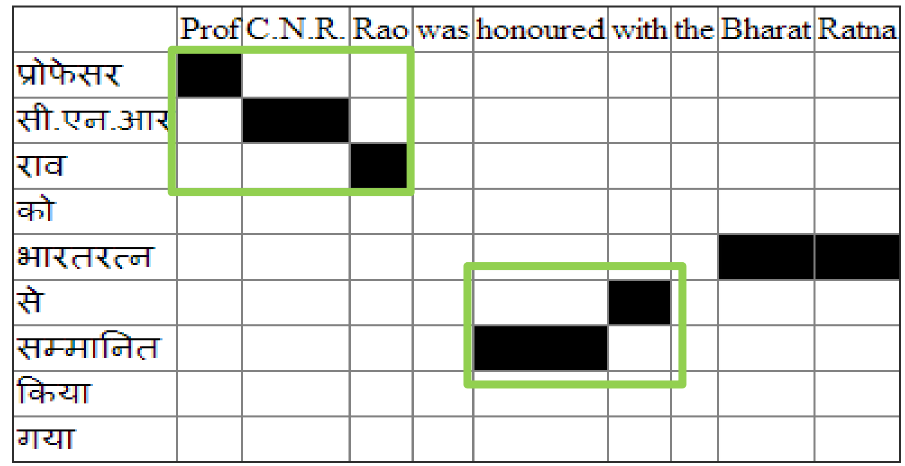
- Our goal is: *arrgmax P(e|f)*
- We decomposed *P(e|f) into P(e) and P(f|e)*
- Considering a source sentence has been segmented into *I* segments, we can further decompose the translation model *P(f|e)* as follows:

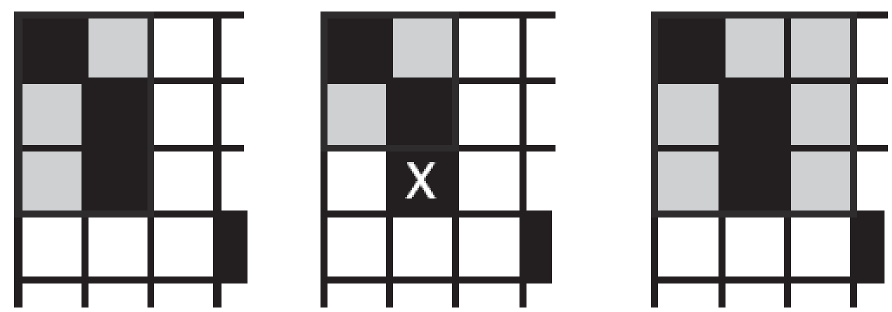
$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(\text{start}_i - \text{end}_{i-1} - 1)$$

- $start_i$: start position in *f* of i^{th} phrase of *e*
- end_i: end position in *f* of i^{th} phrase of *e*
- *ϕ is called as phrase-level translation probability; d* is called as distortion probability

Phrase-based SMT: Phrase Tables

- Involves Structure + Parameter learning
  - Learn the phrase table
  - Learn the phrase-level translation probabilities
- Process:
  - Start with word alignment: reliable input for phrase table learning
  - A consecutive sequence of aligned words constitutes a ‘phrase pair’
- Only ‘consistent’ phrase pairs should be added to the phrase table





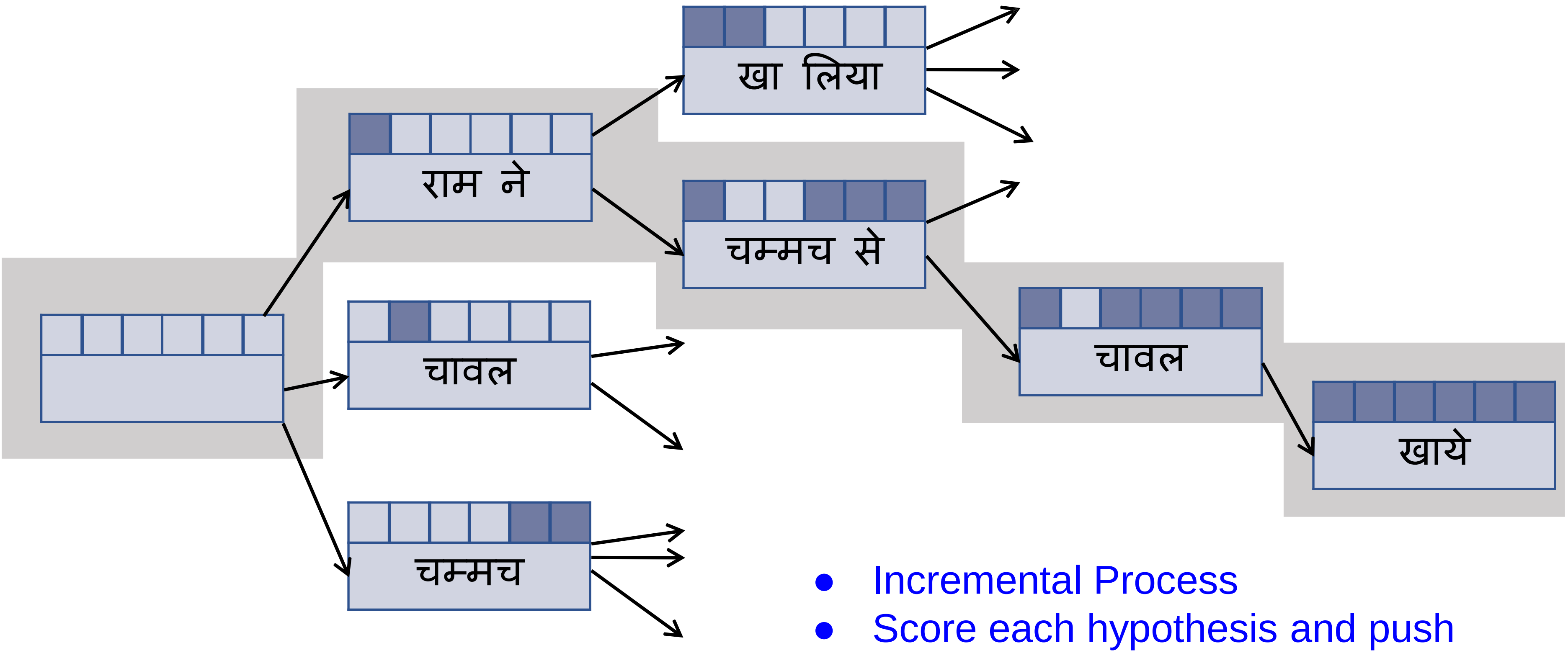
Decoding (1/2)

- Decoding: Searching the best translation in the space of all translations.
- The phrase table may give many options to translate the input sentence leading to multiple word orders.
- Decoding is a NP Complete search problem
  - Needs a heuristic search method

Ram ate rice with the spoon

राम ने चम्मच से चावल खाये

Decoding (2/2)



- Incremental Process
- Score each hypothesis and push them to a bounded priority queue

Paradigms of Machine Translation

Human effort ←

Data hungry →

RBMT → EBMT → SMT → NMT

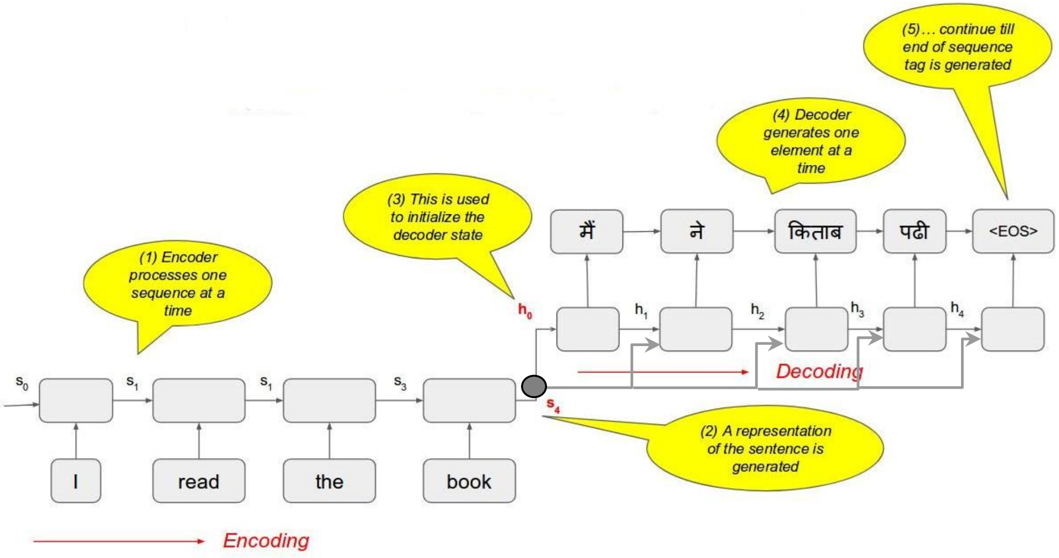
(Rule based machine translation) (Example based machine translation) (Statistical machine translation) (Neural machine translation)

Rules handmade by human expert → Rules by learning from data

What is NMT?

- The task of MT is a sequence-to-sequence problem.
- It uses an encoder-decoder NN architecture with attention mechanism.
- NMT requires large parallel corpus.
- Here, we will discuss RNN-based and Transformer-based encoder-decoder architectures.

Simple RNN-based Encoder-Decoder Architecture



s_4 is the summary vector

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

Summary Vector Representation

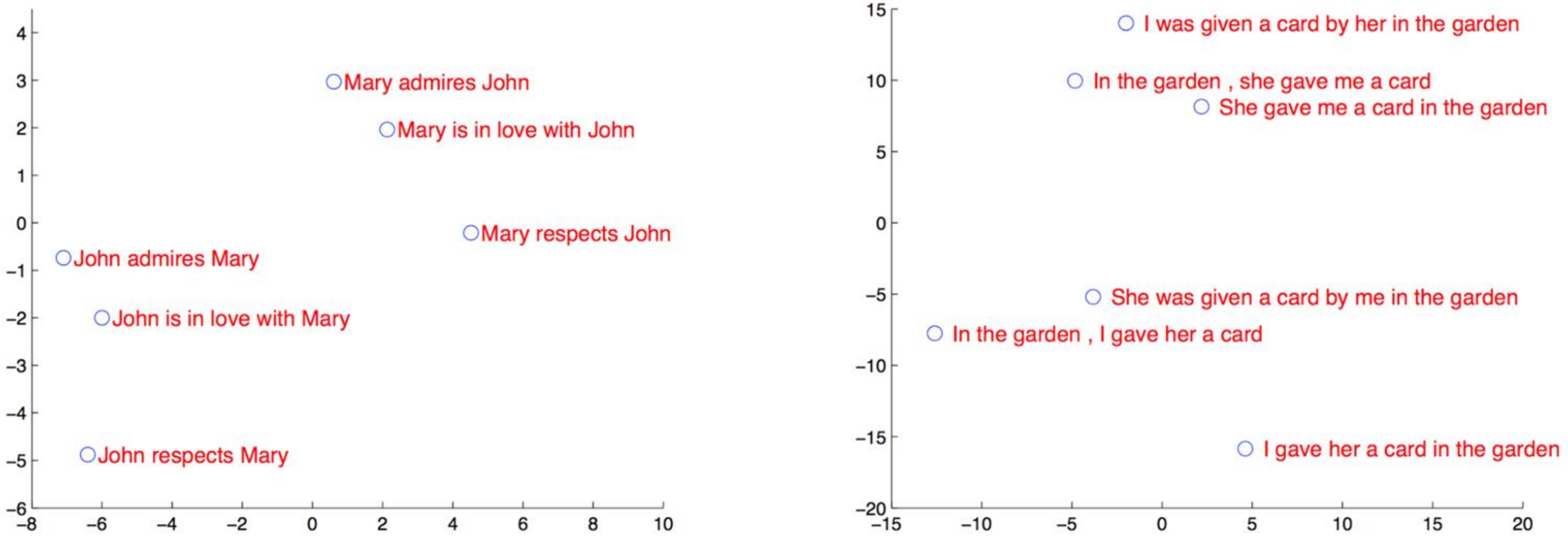


Image source- [9]

Problems with Simple Encode-Decode Paradigm (1/2)

What happens in enc-dec architecture?

1. Encoding transforms the entire sentence into a single vector.
2. Decoding process uses this sentence representation for predicting the output.

Problems:

- Quality of prediction depends upon the quality of sentence embeddings.
- After few time-step, summary vector may lose information of initial words of input sentence.

Problems with Simple Encode-Decode Paradigm (2/2)

Possible Solution:

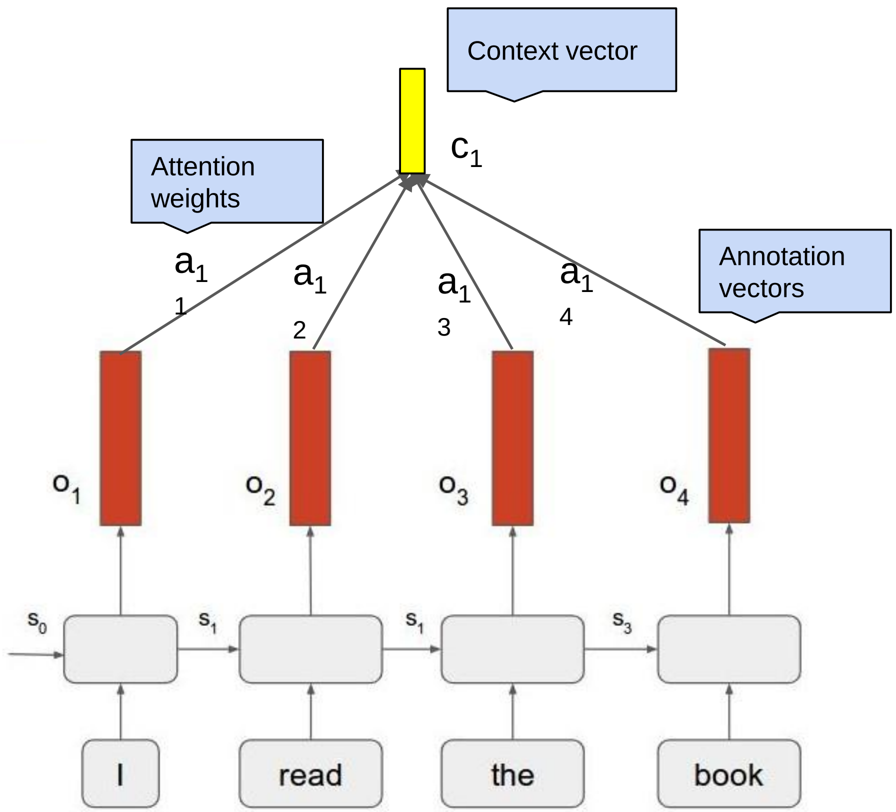
- For prediction at each time step, present the representation of the relevant part of the source sentence only.

the girl goes to school

लड़की स्कूल जाती है

  – Attention-based encoder-decoder

Annotation Vectors and Context Vectors



Attention weights are calculated from alignment scores which are output of another feed-forward NN which is trained jointly.

Attention-based Encoder-Decoder Architecture (1/3)

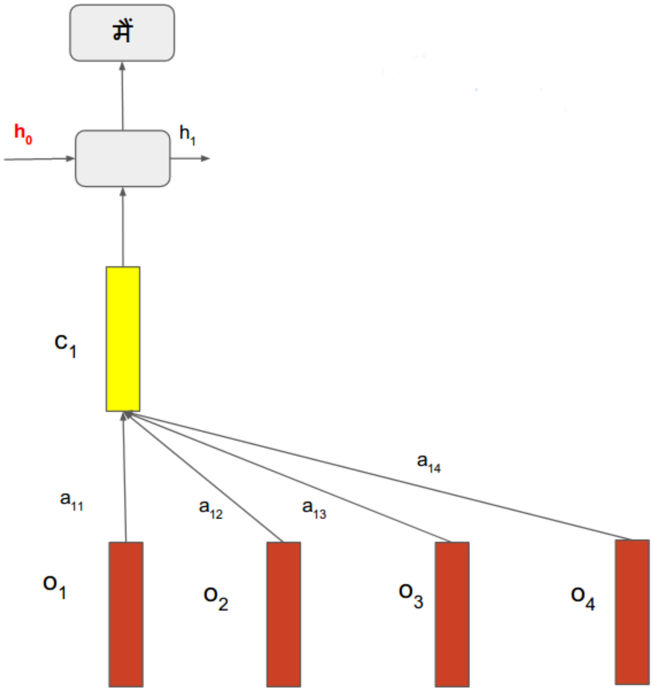


Image source- [http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf]

Attention-based Encoder-Decoder Architecture (2/3)

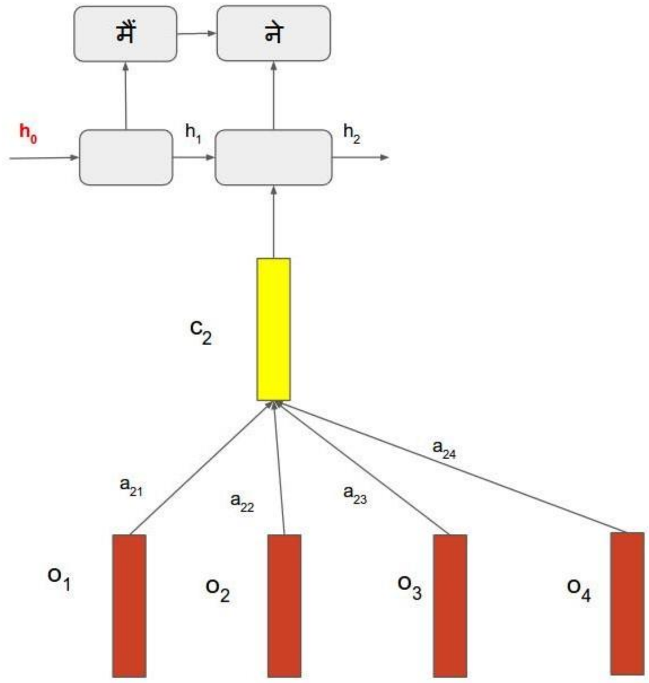


Image source- [http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf]

Attention-based Encoder-Decoder Architecture (3/3)

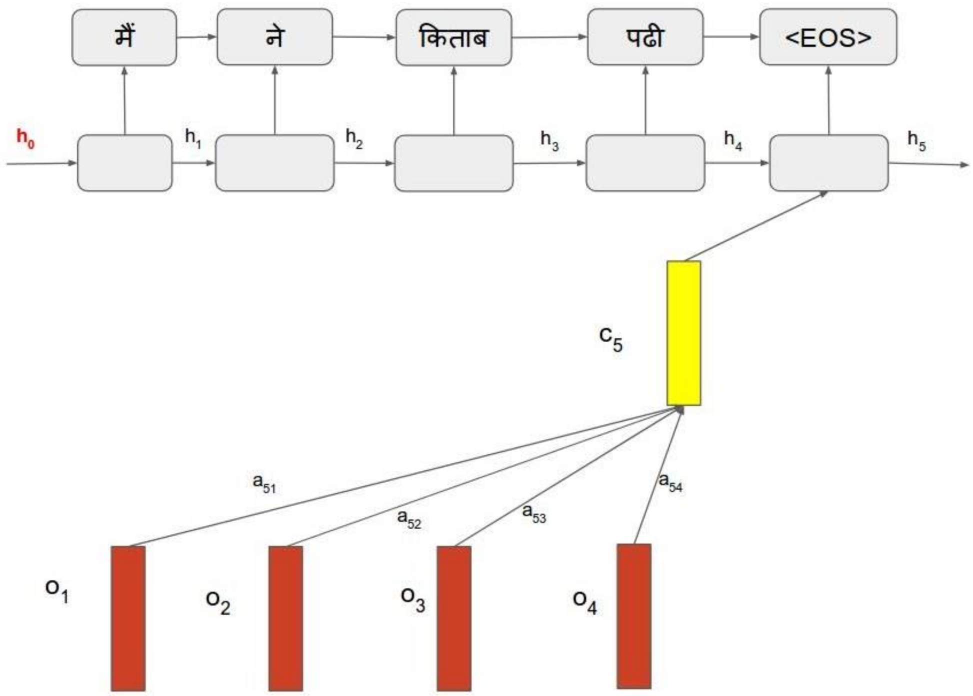


Image source- [http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf]

Main Challenge of MT: Language Divergence

Languages differ in expressing thoughts: Agglutination

- Finnish: “istahtaisinkohan”
- English: "I wonder if I should sit down for a while"

Analysis:

- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix

Kinds of MT Systems

(point of entry from source to the target text)

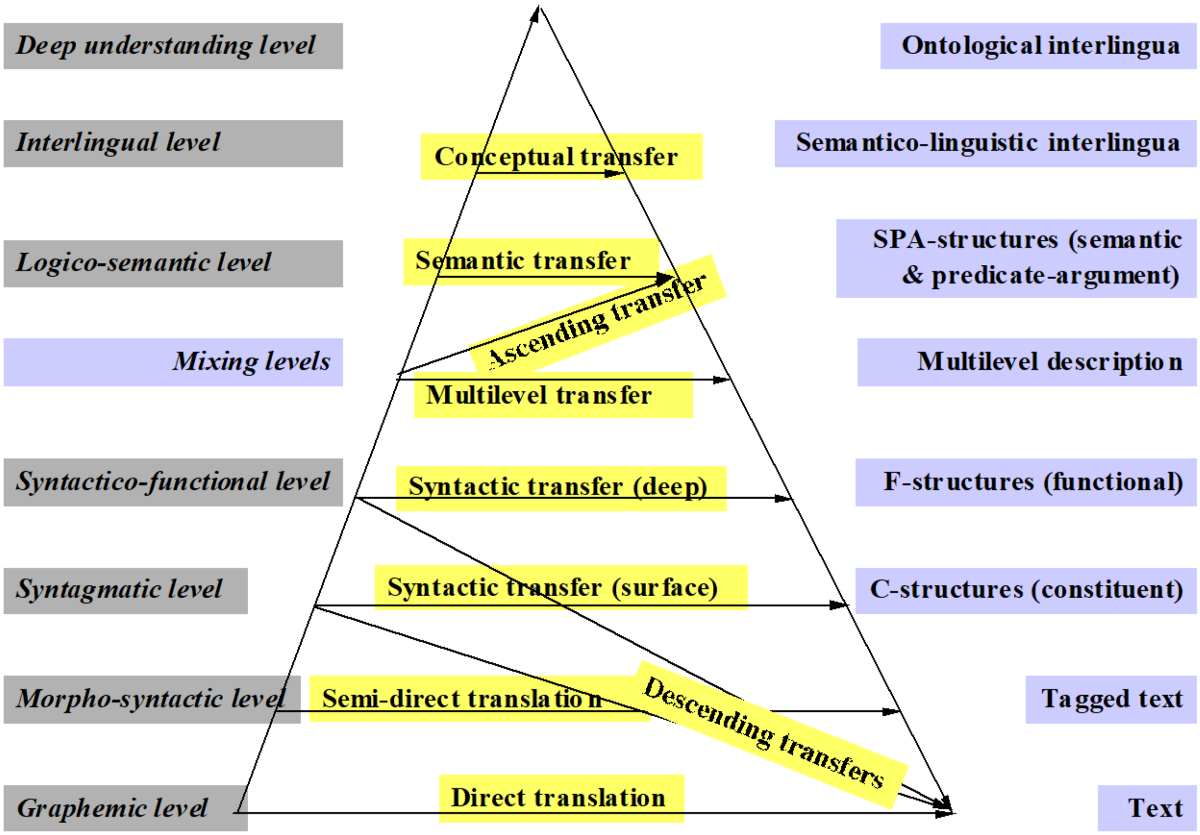


Illustration of transfer SVO→SOV