Machine Translation

Agenda

- What is Machine Translation & why is it interesting?
- Machine Translation Paradigms
- Word Alignment
- Phrase-based SMT
- Extensions to Phrase-based SMT
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

Statistical Machine Translation

Agenda

- **What is Machine Translation & why is it interesting?**
- Machine Translation Paradigms
- Word Alignment
- Phrase-based SMT
- Extensions to Phrase-based SMT
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

What is Machine Translation?

Automatic conversion of text/speech from one natural language to another

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो

bing™ Translator

Google Translate

SDL TRADOS

SYSTRAN beyond language

Machine Translation Usecases

Government

- Administrative requirements
- Education
- Security

Enterprise

- Product manuals
- Customer support

Social

- Travel (signboards, food)
- Entertainment (books, movies, videos)

Translation under the hood

- Cross-lingual Search
- Cross-lingual Summarization
- Building multilingual dictionaries

Any multilingual NLP system will involve some kind of machine translation at some level

Why should you study Machine Translation?

- One of the most challenging problems in Natural Language Processing
- Pushes the boundaries of NLP
- Involves analysis as well as synthesis
- Involves all layers of NLP: morphology, syntax, semantics, pragmatics, discourse
- Theory and techniques in MT are applicable to a wide range of other problems like speech recognition and synthesis

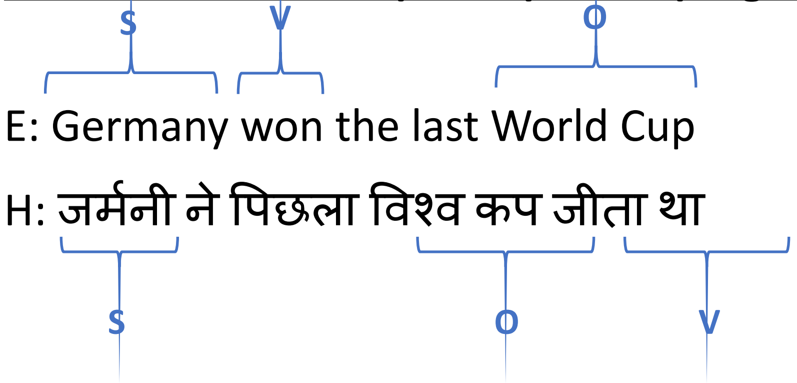
Why is Machine Translation interesting?

Language Divergence □ the great diversity among languages of the world

The central problem of MT is to bridge this language divergence

Language Divergence

Word order: SOV (Hindi), SVO (English), VSO, OSV



E: Germany won the last World Cup

H: जर्मनी ने पिछला विश्व कप जीता था

Free (Hindi) vs rigid (English) word order

पिछला विश्व कप जर्मनी ने जीता था *(correct)*

The last World Cup Germany won *(grammatically incorrect)*

The last World Cup won Germany *(meaning changes)*

Language Divergence

Analytic vs Polysynthetic languages

Analytic (Chinese) □ very few morphemes per word, no inflections

Polysynthetic (Finnish)□ many morphemes per word, no inflections

English: Even if it does not rain

Malayalam: മഴ   പെയ്യുതിലെങ്ങിലും

(rain_noun   shower_verb+not+even_if+then_also)

Inflectional systems [infixing (Arabic), fusional (Hindi), agglutinative (Marathi)]

Arabic

k-t-b: root word

katabtu: I wrote

kattabtu: I had (something) written

kitaab: book

kotub: books

Hindi

Jaaunga (1st per, singular, masculine)

Jaaoge (2nd per)

Jaayega (3rd per, singular, masculine)

Jaayenge (3rd per, plural)

Marathi

कपाटावरील: कपाट + वर + ईल

(the one over the cupboard)

दारावरील: दार + वर + ईल

(the one over the door)

दारामागील: दार + मागे + ईल

(the one behind the door)

Language Divergence

Different ways of expressing same concept

water □ पानी, जल, नीर

Language registers

Formal: आप बैठिये Informal: तू बैठ

Standard : मुझे डोसा चाहिए Dakhini: मेरे को डोसा होना

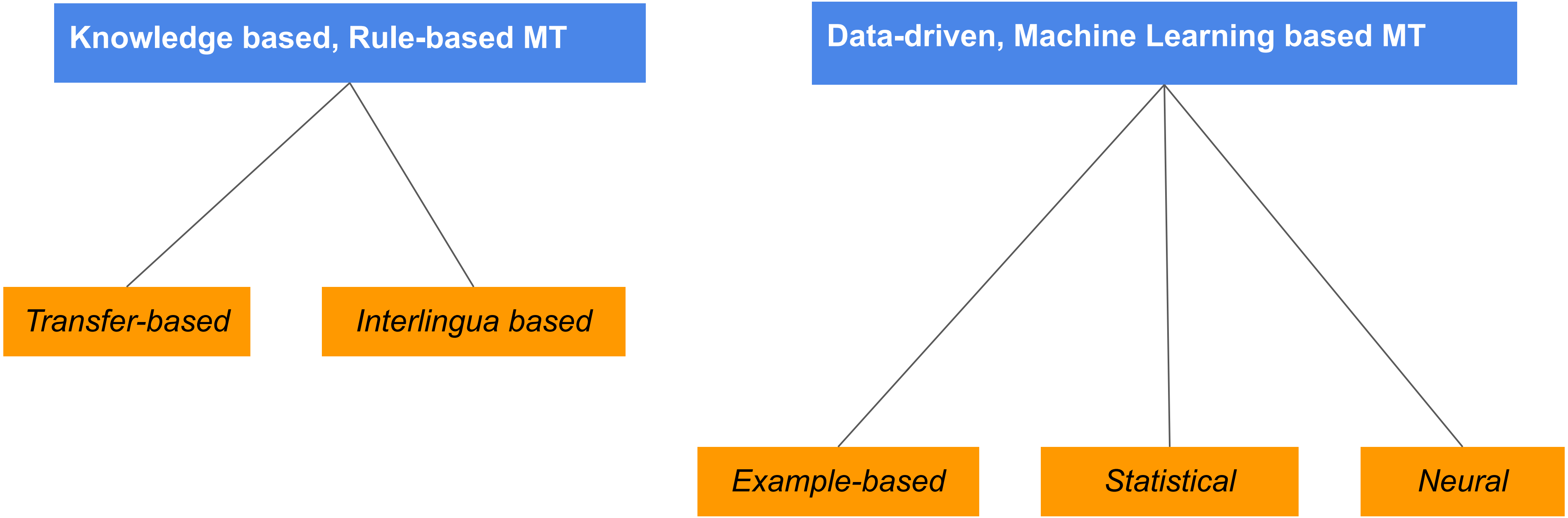
Why is Machine Translation difficult?

- **Ambiguity**
  - Same word, multiple meanings: मंत्री (minister or chess piece)
  - Same meaning, multiple words: जल, पानी, नीर (water)
- **Word Order**
  - Underlying deeper syntactic structure
  - Phrase structure grammar?
  - Computationally intensive
- **Morphological Richness**
  - Identifying basic units of words

Agenda

- What is Machine Translation & why is it interesting?
- **Machine Translation Paradigms**
- Word Alignment
- Phrase-based SMT
- Extensions to Phrase-based SMT
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

Approaches to build MT systems



Knowledge based, Rule-based MT

Transfer-based

Interlingua based

Data-driven, Machine Learning based MT

Example-based

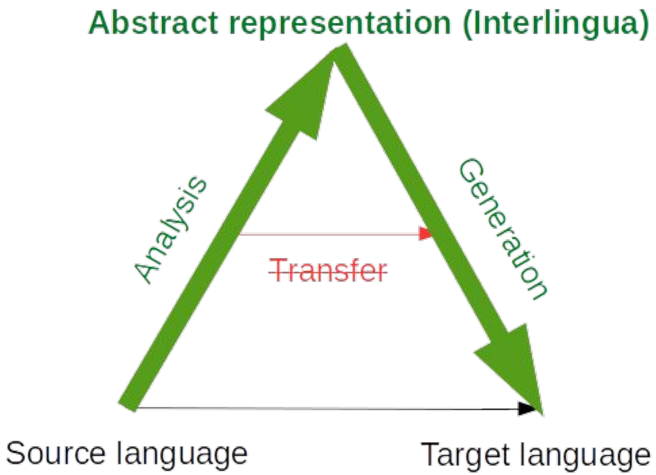
Statistical

Neural

Rule-based MT

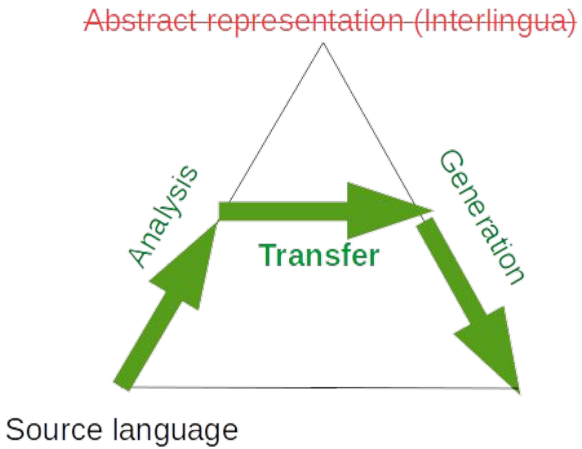
- Rules are written by ***linguistic experts*** to analyze the source, generate an intermediate representation, and generate the target sentence
- Depending on the depth of analysis: interlingua or transfer-based MT

Interlingua based MT



Deep analysis, complete disambiguation and language independent representation

Transfer based MT



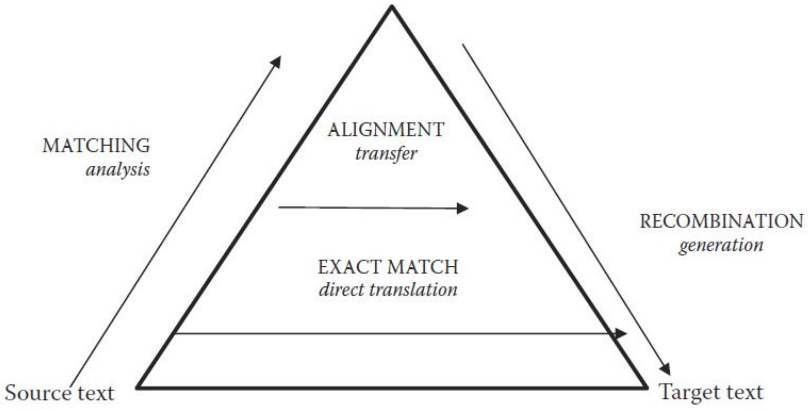
Partial analysis, partial disambiguation and a bridge intermediate representation

Problems with rule-based MT

- Required linguistic expertise to develop systems
- Maintenance of system is difficult
- Difficult to handle ambiguity
- Scaling to a large number of language pairs is not easy

Example-based MT

Translation by analogy ⇒ match parts of sentences to known translations and then combine



Input: *He buys a book on international politics*

1. **Phrase fragment matching: (*data-driven*)**

   he buys
   a book
   international politics

2. **Translation of segments: (*data-driven*)**

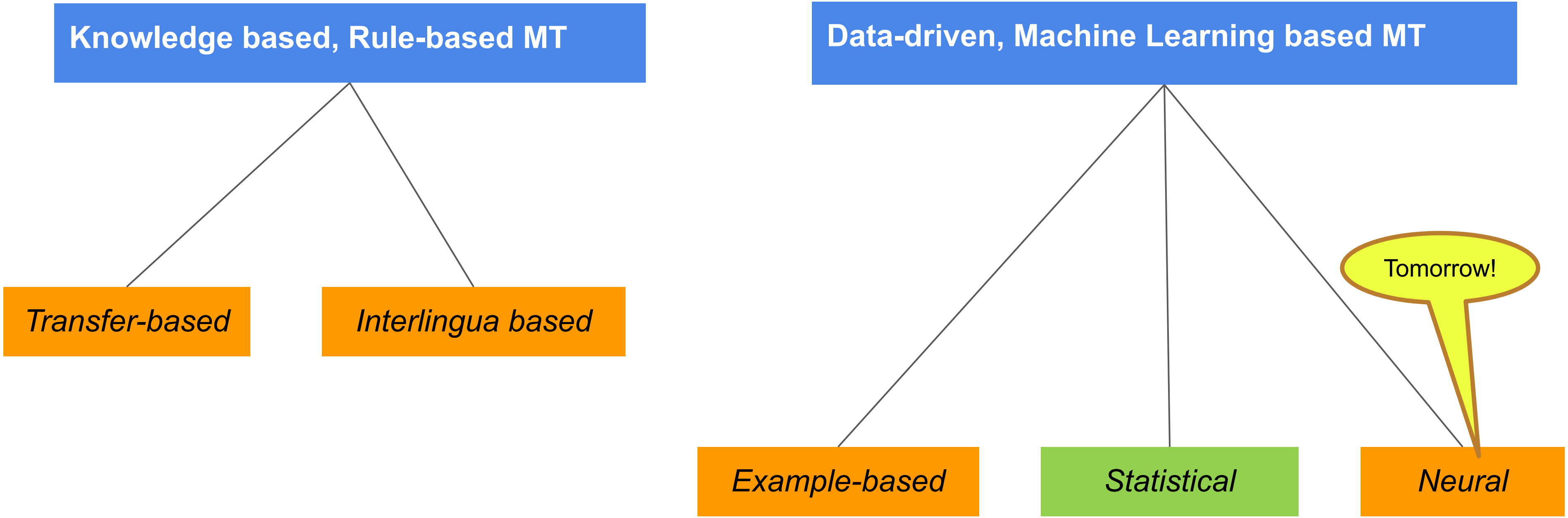
   वह खरीदता है
   एक किताब
   अंतर राष्ट्रीय राजनीति

3. **Recombination:** *(human crafted rules/templates)*

   वह अंतर राष्ट्रीय राजनीति पर एक किताब खरीदता है

- *Partly rule-based, partly data-driven.*
- *Good methods for matching and large corpora did not exist when proposed*

Approaches to build MT systems



Knowledge based, Rule-based MT

Transfer-based

Interlingua based

Data-driven, Machine Learning based MT

Example-based

Statistical

Neural

Tomorrow!

Statistical Machine Translation

A Probabilistic Formalism

Let’s formalize the translation process

*We will model translation using a **probabilistic model**. Why?*

- *We would like to have a measure of confidence for the translations we learn*
- *We would like to model uncertainty in translation*

E: target language
F: source language

e: source language sentence
f : target language sentence

Best translation

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

How do we **model** this quantity?

***Model**: a simplified and idealized understanding of a physical process*

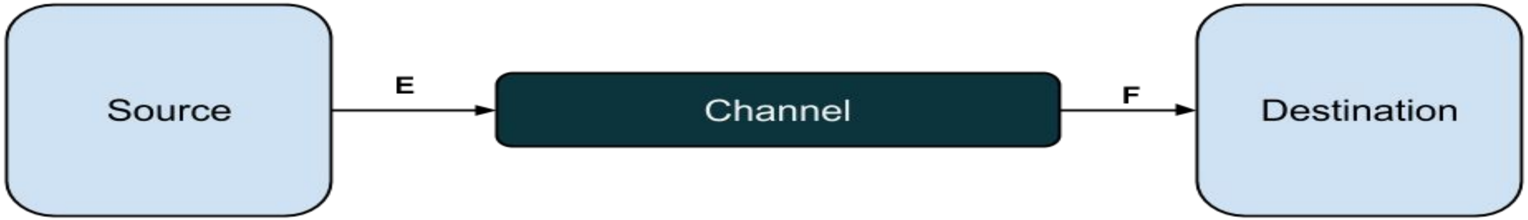
We explain translation using the Noisy Channel Model

A very general framework for many NLP problems

Generate target sentence

Channel corrupts the target

Source sentence is a corruption of the target sentence



Language Model (LM)
P(e)

Captures fluency

Translation Model (TM)
P(f|e)

Captures fidelity

$$P(e|f) = P(e) \times P(f|e)$$

Translation is the process of recovering the original signal given the corrupted signal

Why use this counter-intuitive way of explaining translation?

- Makes it easier to mathematically represent translation and learn probabilities

We have already seen how to learn n-gram language models

Let’s see how to learn the translation model → $P(\boldsymbol{f}|\boldsymbol{e})$

To learn sentence translation probabilities,

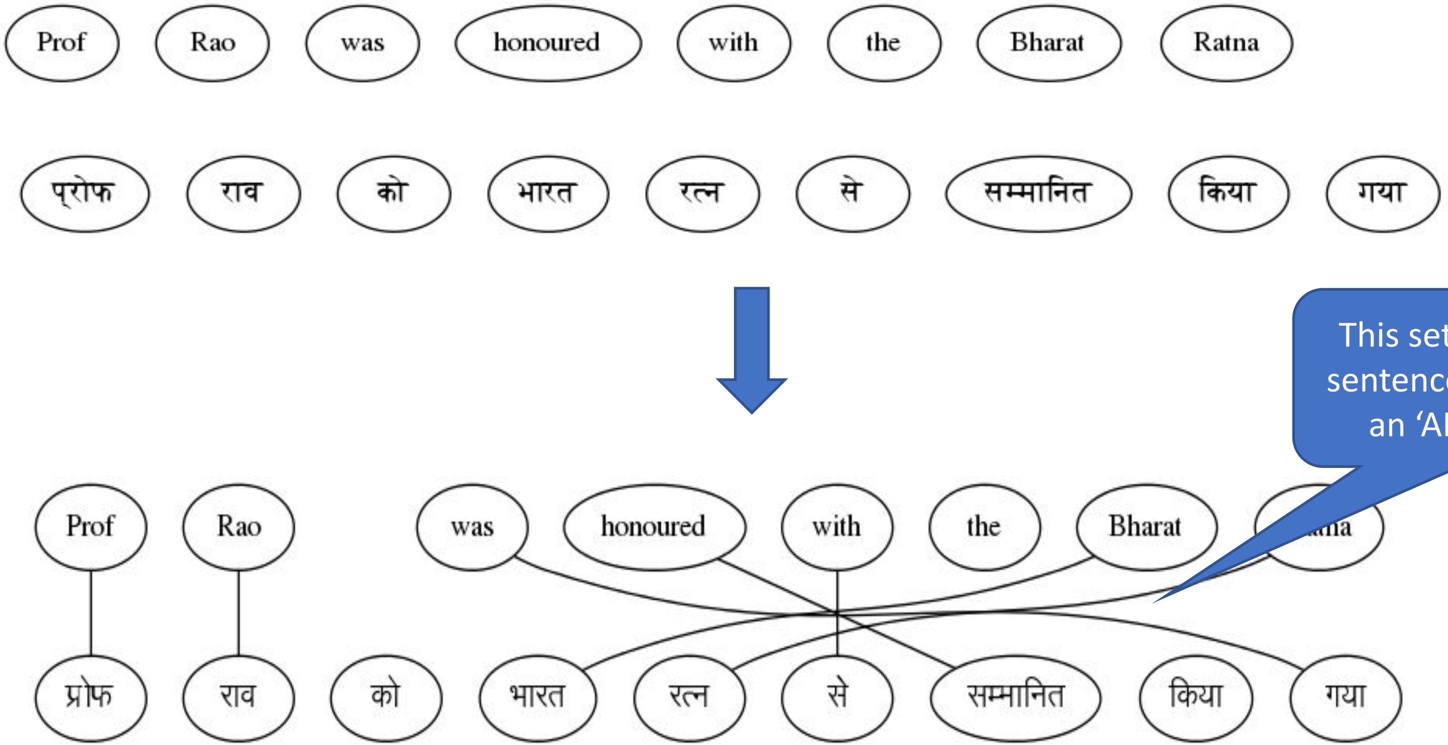
□ we first need to learn word-level translation probabilities

That is the task of word alignment

Agenda

- What is Machine Translation & why is it interesting?
- Machine Translation Paradigms
- **<u>Word Alignment</u>**
- Phrase-based SMT
- Extensions to Phrase-based SMT
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

Given a parallel sentence pair, find word level correspondences



Prof Rao was honoured with the Bharat Ratna

प्रोफ राव को भारत रत्न से सम्मानित किया गया

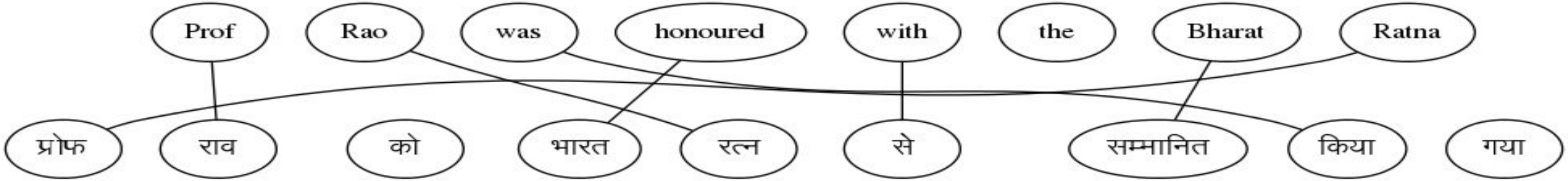
Prof Rao was honoured with the Bharat Ratna

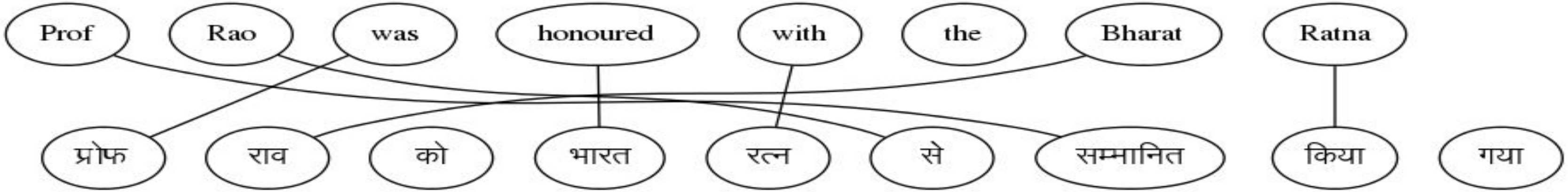
प्रोफ राव को भारत रत्न से सम्मानित किया गया

This set of links for a sentence pair is called an ‘ALIGNMENT’

But there are multiple possible alignments

Sentence 1

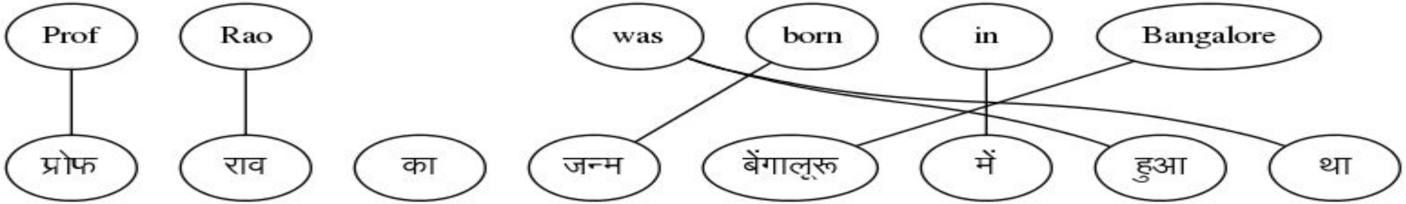


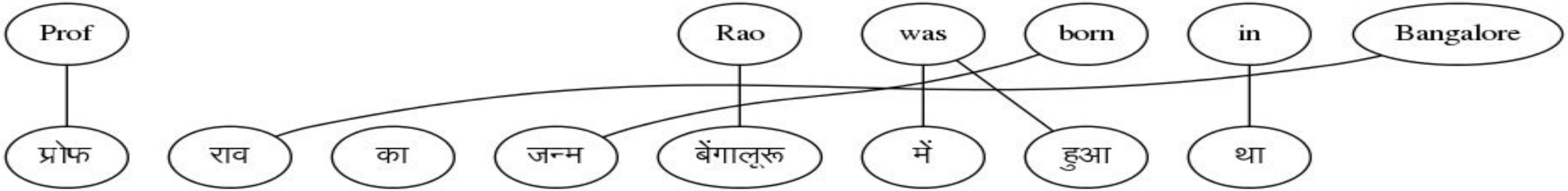


With one sentence pair, we cannot find the correct alignment

Can we find alignments if we have multiple sentence pairs?

Sentence 2





Yes, let’s see how to do that …

Parallel Corpus

A boy is sitting in the kitchen	एक लडका रसोई म़े 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 पढ रही है
A woman is sitting in a red car	एक औरत एक काले कार मे बैठी है

Parallel Corpus

A boy is **sitting** in the kitchen	एक लडका रसोई मे़ **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 पढ रही है
A woman is **sitting** in a red car	एक औरत एक काले कार मे **बैठा** है

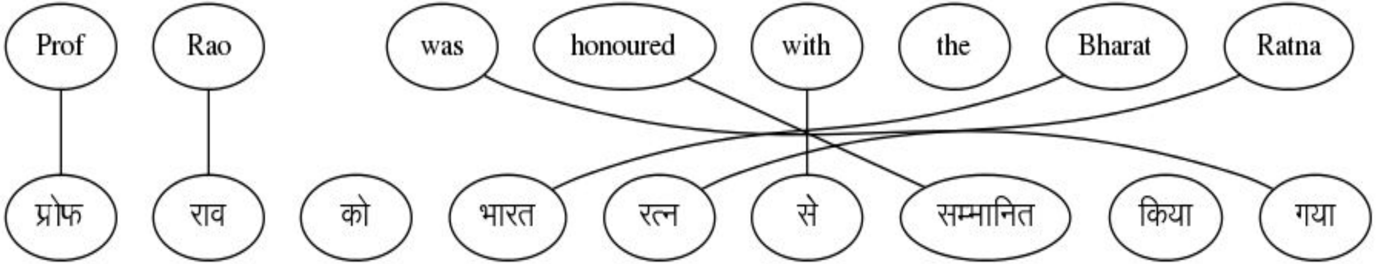
Key Idea

Co-occurrence of translated words

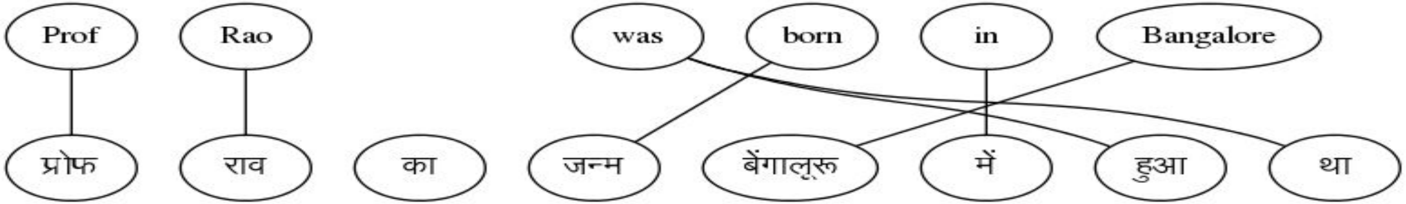
Words which occur together in the parallel sentence are likely to be translations (higher P(f|e))

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

Sentence 1



Sentence 2



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

$$P(Prof|\text{प्रोफ}) = \frac{2}{2}$$

$\#(a,b)$: *number of times word a is aligned to word b*

But, we can find the best alignment only if we know the word translation probabilities

The best alignment is the one that maximizes the sentence translation probability

$$P(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e}) = P(a)\prod_{i=1}^{i=m} P\left(f_i\middle|e_{a_i}\right) \quad\Longrightarrow\quad \boldsymbol{a}^* = \underset{\boldsymbol{a}}{\operatorname{argmax}} \prod_{i=1}^{i=m} P(f_i|e_{a_i})$$

This is a chicken and egg problem! How do we solve this?

We can solve this problem using a two-step, iterative process

Start with random values for word translation probabilities

Step 1: Estimate alignment probabilities using word translation probabilities

Step 2: Re-estimate word translation probabilities

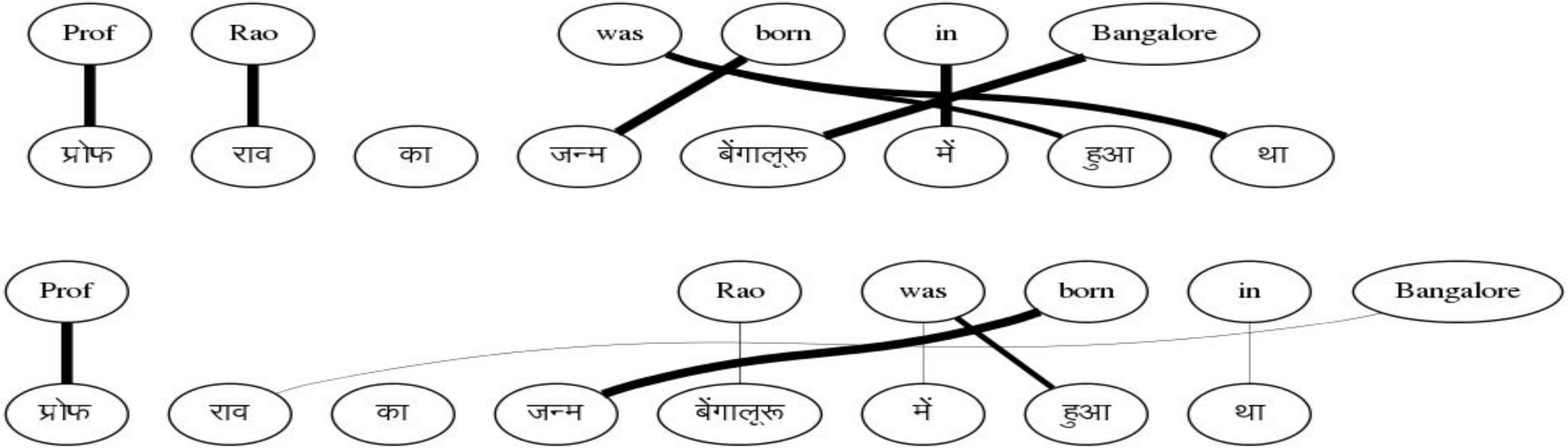
- *We don't know the best alignment*
- *So, we consider all alignments while estimating word translation probabilities*
  - *Instead of taking only the best alignment, we consider all alignments and weigh the word alignments with the alignment probabilities*

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

Repeat Steps (1) and (2) till the parameters converge

At the end of the process …

Sentence 2



Is the algorithm guaranteed to converge?

That’s the nice part □ it is guaranteed to converge

This is an example of the well known Expectation-Maximization Algorithm

However, the problem is highly non-convex

Will lead to local minima

Good modelling assumptions necessary to ensure a good solution

Agenda

- What is Machine Translation & why is it interesting?
- Machine Translation Paradigms
- Word Alignment
- **Phrase-based SMT**
- Extensions to Phrase-based SMT
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

What is PB-SMT?

Why stop at learning word correspondences?

KEY IDEA □ Use “Phrase” (Sequence of Words) as the basic translation unit

Note: the term ‘phrase’ is not used in a linguistic sense

The Prime Minister of India	भारत के प्रधान मंत्री bhArata ke pradhAna maMtrI India of Prime Minister
is running fast	तेज भाग रहा है teja bhAg rahA hai fast run -continuous is
honoured with	से सम्मानित किया se sammanita kiyA with honoured did
Rahul lost the match	राहुल मुकाबला हार गया rAhula mukAbalA hAra gayA Rahul match lost

Parallel Corpus

A boy is **sitting** in the kitchen	एक लडका रसोई मे़ **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 पढ रही है
A woman is **sitting** in a red car	एक औरत एक काले कार मे **बैठा** है

Benefits of PB-SMT

Local Reordering □ Intra-phrase re-ordering can be memorized

The Prime Minister of India	भारत के प्रधान मंत्री bhaarat ke pradhaan maMtrI India of Prime Minister

Sense disambiguation based on local context □ Neighbouring words help make the choice

heads towards Pune	पुणे की ओर जा रहे है pune ki or jaa rahe hai Pune towards go –continuous is
heads the committee	समिति की अध्यक्षता करते है Samiti kii adhyakshata karte hai committee of leading -verbalizer is

Benefits of PB-SMT (2)

Handling institutionalized expressions

- Institutionalized expressions, idioms can be learnt as a single unit

hung assembly	त्रिशंकु विधानसभा trishanku vidhaansabha
Home Minister	गृह मंत्री gruh mantrii
Exit poll	चुनाव बाद सर्वेक्षण chunav baad sarvekshana

- Improved Fluency
  - The phrases can be arbitrarily long (even entire sentences)

Mathematical Model

<u>Let’s revisit the decision rule for SMT model</u>

$$\begin{aligned}\mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{\text{LM}}(\mathbf{e})\end{aligned}$$

<u>Let’s revisit the translation model *p*(**f**|**e**)</u>

- Source sentence can be segmented in $\mathbf{I}$ phrases
- Then, *p*(**f**|**e**) can be decomposed as:

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

Distortion probability

Phrase Translation Probability

start_i :start position in **f** of ith phrase of **e**
end_i :end position in **f** of ith phrase of **e**

Learning The Phrase Translation Model

Involves Structure + Parameter Learning:

- Learn the **Phrase Table**: the central data structure in PB-SMT

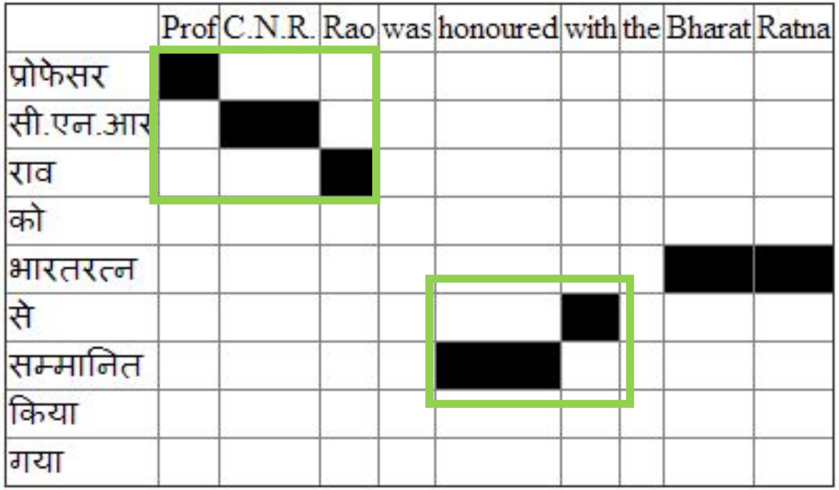
The Prime Minister of India	भारत के प्रधान मंत्री
is running fast	तेज भाग रहा है
the boy with the telescope	दूरबीन से लड़के को
Rahul lost the match	राहुल मुकाबला हार गया

- Learn the **Phrase Translation Probabilities**

Prime Minister of India	भारत के प्रधान मंत्री India of Prime Minister	0.75
Prime Minister of India	भारत के भूतपूर्व प्रधान मंत्री India of former Prime Minister	0.02
Prime Minister of India	प्रधान मंत्री Prime Minister	0.23

Learning Phrase Tables from Word Alignments

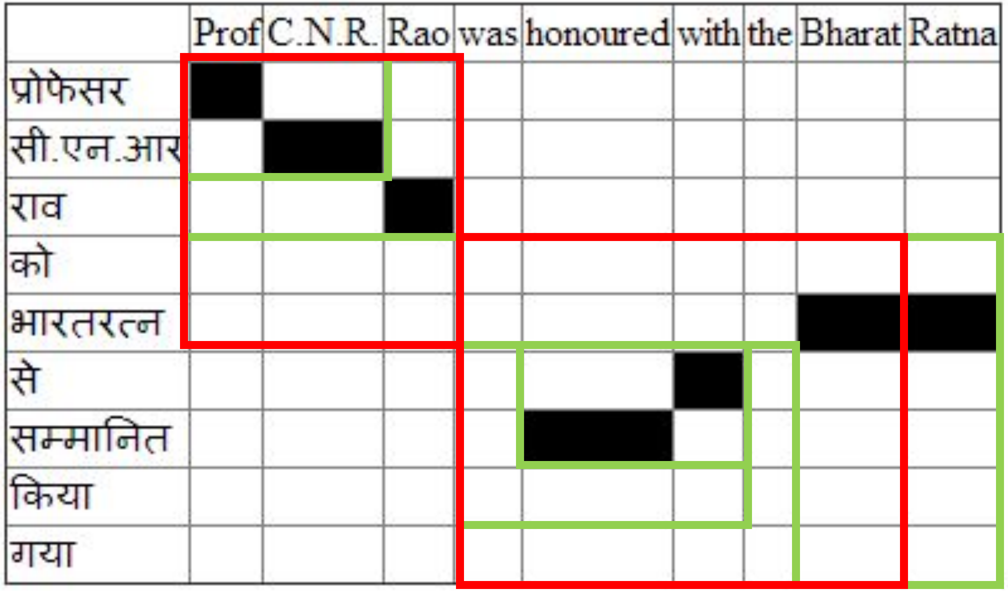
- Start with word alignments
- Word Alignment : reliable input for phrase table learning
  - high accuracy reported for many language pairs
- Central Idea: A consecutive sequence of aligned words constitutes a “phrase pair”



	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न								■	■
से						■			
सम्मानित					■				
किया									
गया									

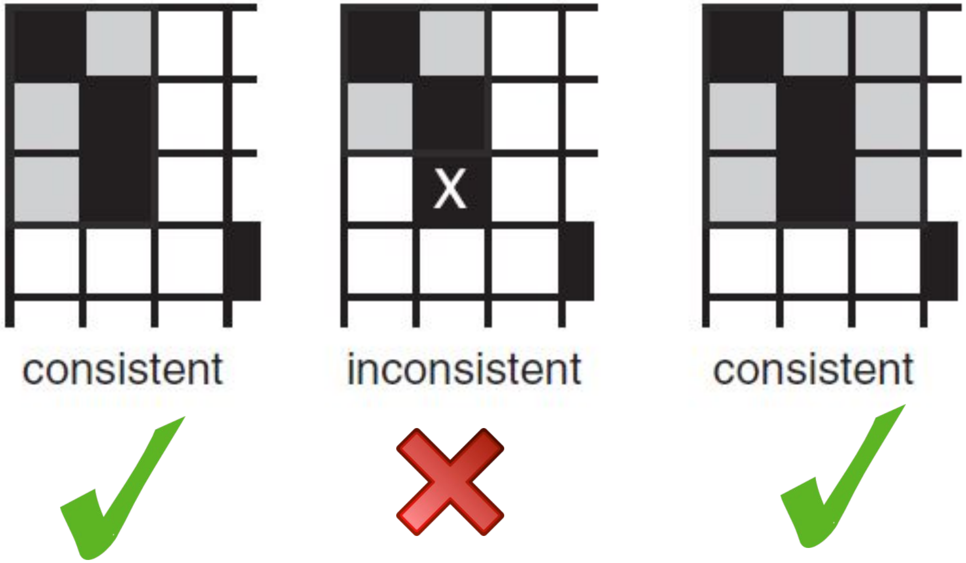
Which phrase pairs to include in the phrase table?

Extracting Phrase Pairs



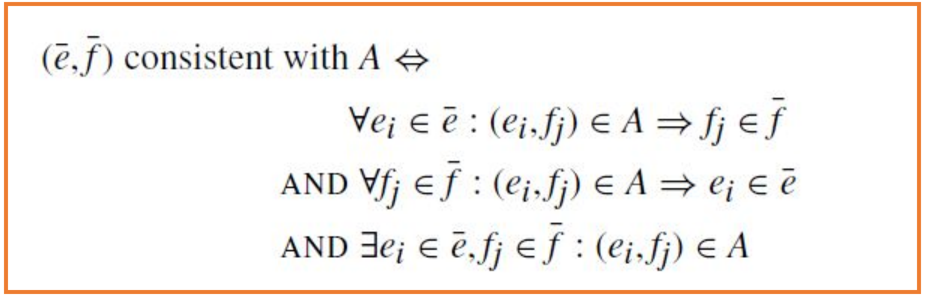
	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न								■	■
से						■			
सम्मानित					■				
किया									
गया									

Phrase Pairs “consistent” with word alignment



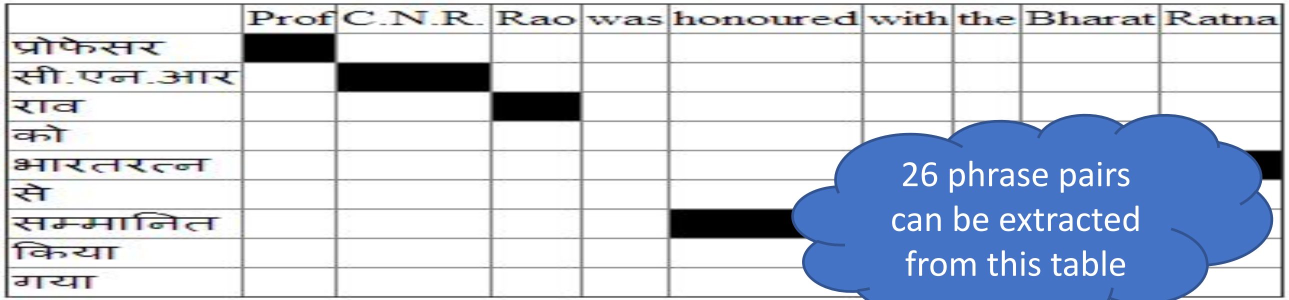
Source: SMT, Phillip Koehn

Phrase Pairs “consistent” with word alignment



$$
\begin{aligned}
(\bar{e},\bar{f}) \text{ consistent with } A \Leftrightarrow & \\
& \forall e_i \in \bar{e} : (e_i,f_j) \in A \Rightarrow f_j \in \bar{f} \\
\text{AND } & \forall f_j \in \bar{f} : (e_i,f_j) \in A \Rightarrow e_i \in \bar{e} \\
\text{AND } & \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i,f_j) \in A
\end{aligned}
$$

Examples



	Prof	C.N.R.	Rao	was	honoured	with	the	Bharat	Ratna
प्रोफेसर	■								
सी.एन.आर		■							
राव			■						
को									
भारतरत्न									■
से									
सम्मानित					■				
किया									
गया									

26 phrase pairs can be extracted from this table

Professor CNR	प्रोफेसर सी.एन.आर
Professor CNR Rao	प्रोफेसर सी.एन.आर राव
Professor CNR Rao was	प्रोफेसर सी.एन.आर राव
Professor CNR Rao was	प्रोफेसर सी.एन.आर राव को
honoured with the Bharat Ratna	भारतरत्न से सम्मानित
honoured with the Bharat Ratna	भारतरत्न से सम्मानित किया
honoured with the Bharat Ratna	भारतरत्न से सम्मानित किया गया
honoured with the Bharat Ratna	को भारतरत्न से सम्मानित किया गया

Computing Phrase Translation Probabilities

- Estimated from the relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e},\bar{f})}{\sum_{\bar{f_i}} \text{count}(\bar{e},\bar{f_i})}$$

Prime Minister of India	भारत के प्रधान मंत्री India of Prime Minister	***0.75***
Prime Minister of India	भारत के भूतपूर्व प्रधान मंत्री India of former Prime Minister	***0.02***
Prime Minister of India	प्रधान मंत्री Prime Minister	***0.23***

Generative vs. Discriminative models in ML

Generative Model

- Noisy channel model of translation from sentence f to sentence e.
- Task is to recover e from noisy f.

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f}|\mathbf{e})$$

P(f|e): Translation model, addresses adequacy

P(e): Language model, addresses fluency

- Joint modeling of entire parameter space
- The generative story is too simplistic, not reflective of translation process

Discriminative Model

- Maximum Entropy based model, incorporating arbitrary features

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{e} \exp \sum_{i} \lambda_i h_i(f, e)$$

- h_i - features functions, λ_i are feature weights
- No need to model source, reduces parameter space
- Arbitrary features can better capture translation process
- Why exponential function form? –maximizing entropy w.r.t data constraints

Discriminative Training of PB-SMT

- Directly model the posterior probability p**(e|f)**
- Use the Maximum Entropy framework

$$P(\mathbf{e}|\mathbf{f}) = \exp\left(\sum_i \lambda_i h_i(f_1^I, e_1^J)\right)$$

$$e^* = \arg\max_{e_i} \sum_i \lambda_i h_i(f_1^I, e_1^J)$$

  - h_i**(f,e)** are feature functions , λ_i's are feature weights
- Benefits:
  - Can add arbitrary features to score the translations
  - Can assign different weight for each features
  - Assumptions of generative model may be incorrect

Generative Model as a special case

Generative model

$$\begin{aligned}\mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{\text{LM}}(\mathbf{e})\end{aligned}$$

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

Feature function mappings for corresponding discriminative model

$$h_1 = \prod_{i=1}^{I} \phi(\bar{f}_i, \bar{e}_i) \quad , \quad \lambda_1 = 1$$

translation model

distortion model

$$h_2 = \prod_{i=1}^{I} d(start_i - end_{i-1} - 1) \quad , \quad \lambda_2 = 1$$

$$h_3 = p_{\text{LM}}(\mathbf{e}) \quad , \quad \lambda_3 = 1$$

language model

More features for PB-SMT

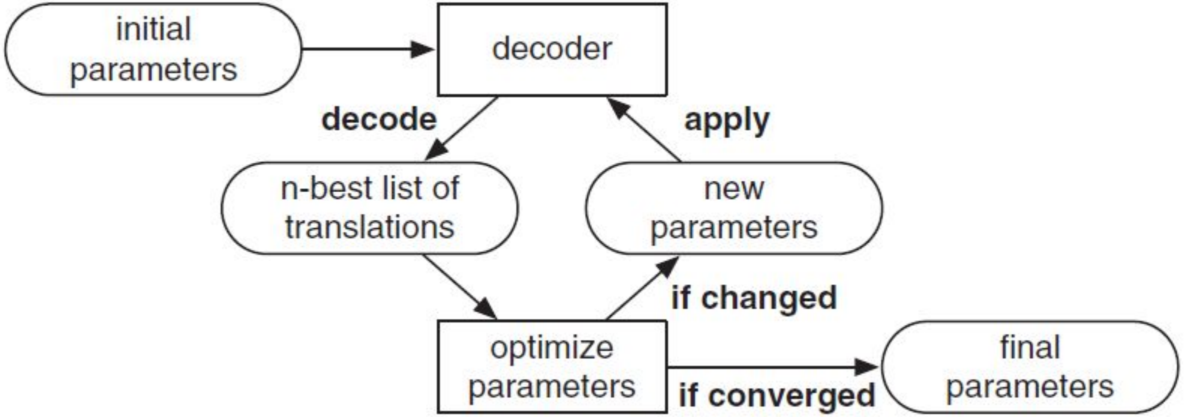
- Inverse phrase translation probability ($\phi(\bar{f}|\bar{e})$)
- Lexical Weighting

$$\text{lex}(\bar{e}|\bar{f}, a) = \prod_{i=1}^{\text{length}(\bar{e})} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall(i,j) \in a} w(e_i|f_j)$$

  - *a:* alignment between words in phrase pair (ē, f)
  - *w(x|y):* word translation probability
- Inverse Lexical Weighting
  - Same as above, in the other direction

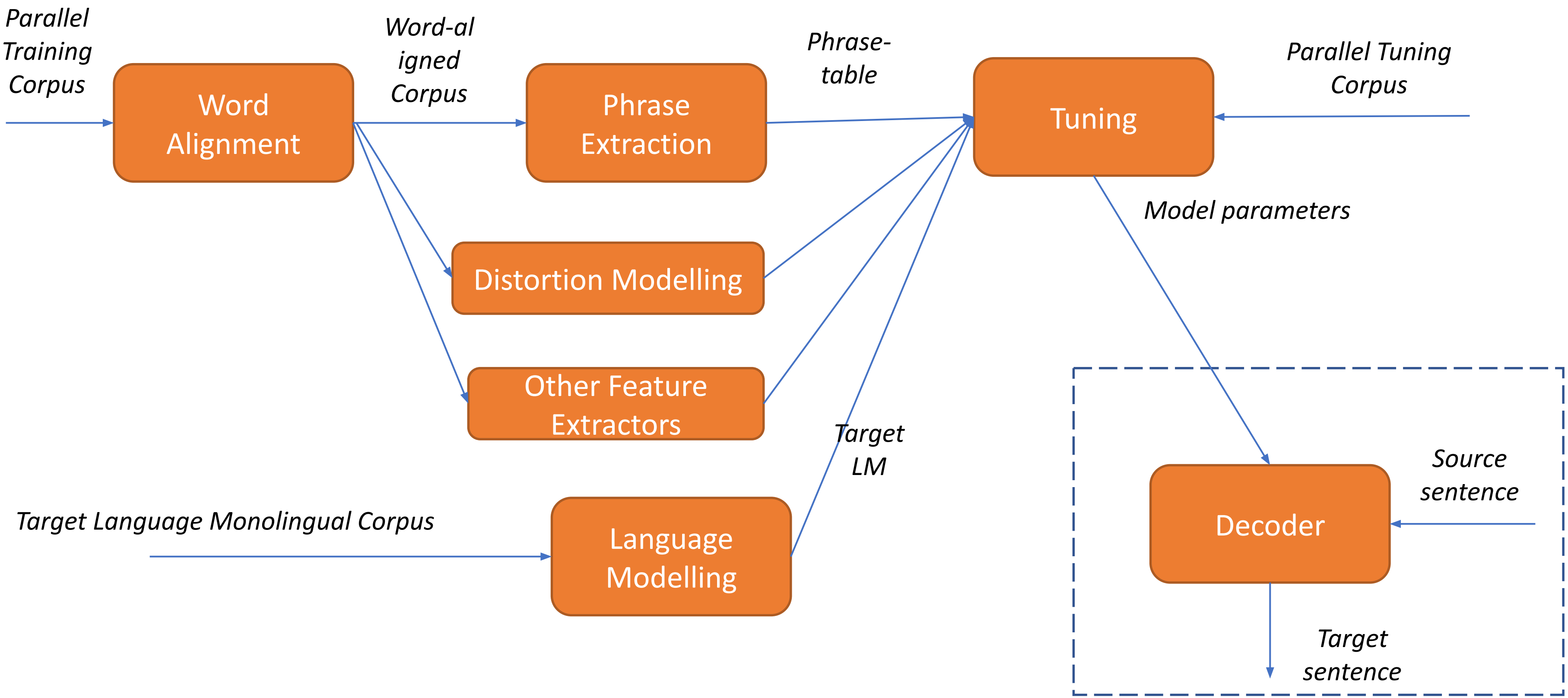
Tuning

- Learning feature weights from data – λ_i
- Minimum Error Rate Training (MERT)
- Search for weights which minimize the translation error on a held-out set (tuning set)
  - Translation error metric : (1 – *BLEU)*



Source: SMT, Phillip Koehn

Typical SMT Pipeline

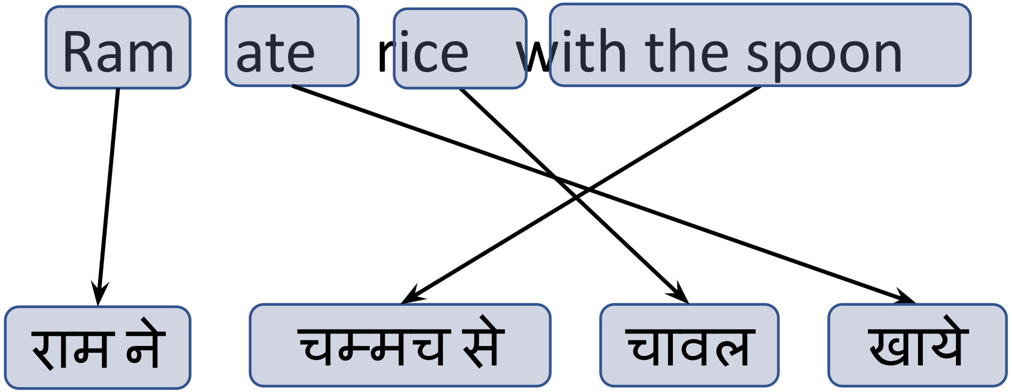


Decoding

Searching for the best translations in the space of all translations

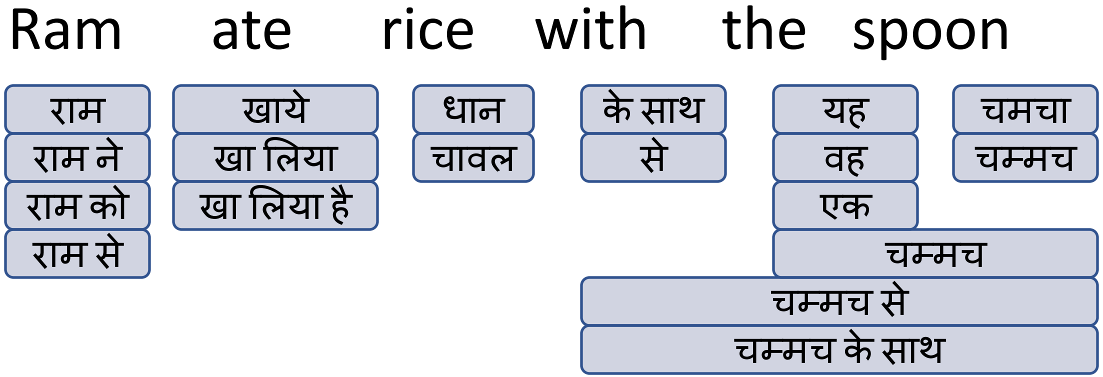
$$e^* = \arg\max_{e_i} \sum_i \lambda_i h_i(f_1^I, e_1^J)$$

An Example of Translation



Reality

- We picked the phrase translation that made sense to us
- The computer has less intuition
- Phrase table may give many options to translate the input sentence



Ram ate rice with the spoon

राम
राम ने
राम को
राम से

खाये
खा लिया
खा लिया है

धान
चावल

के साथ
से

यह
वह
एक

चमचा
चम्मच

चम्मच

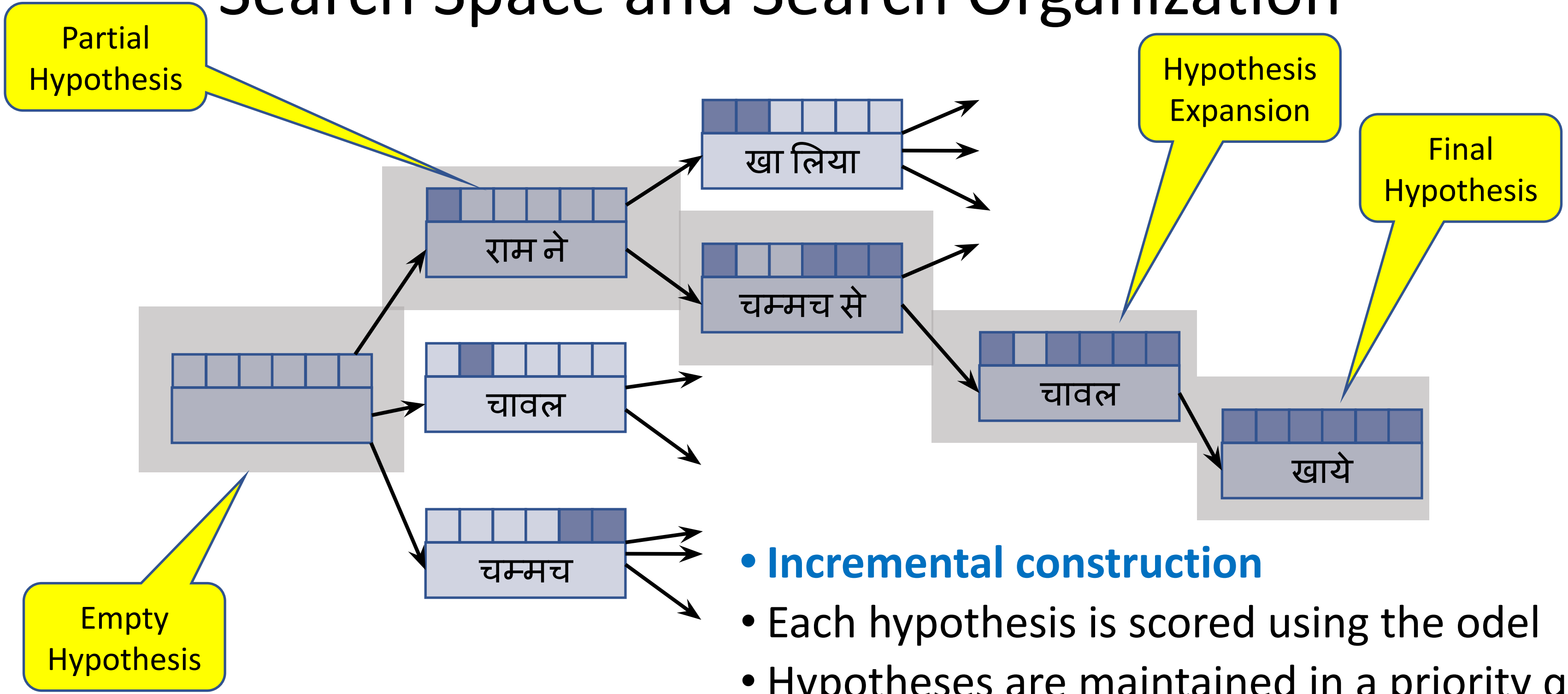
चम्मच से

चम्मच के साथ

What is the challenge in decoding?

- The task of decoding in machine translation is to find the best scoring translation according to translation models
- Hard problem, since there is a exponential number of choices, given a specific input sentence
- Shown as an NP complete problem
- Need to come up with heuristic search methods
- No guarantee of finding the best translation

Search Space and Search Organization



- **Incremental construction**
- Each hypothesis is scored using the odel
- Hypotheses are maintained in a priority queue
- Limit to the reordering window for efficiency

Agenda

- What is Machine Translation & why is it interesting?
- Machine Translation Paradigms
- Word Alignment
- Phrase-based SMT
- **Extensions to Phrase-based SMT**
  - Addressing Word-order Divergence
  - Addressing Morphological Divergence
  - Handling Named Entities
- Syntax-based SMT
- Machine Translation Evaluation
- Summary

We have looked at a basic phrase-based SMT system

This system can learn word and phrase translations from parallel corpora

But many important linguistic phenomena need to be handled

- **Divergent Word Order**
- Rich morphology
- Named Entities and Out-of-Vocabulary words

Getting word order right

Phrase based MT is not good at learning word ordering

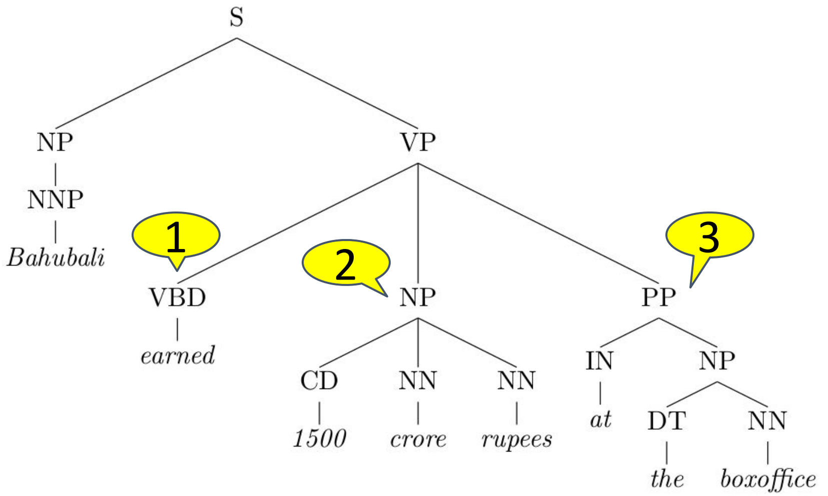
Solution: Let’s help PB-SMT with some preprocessing of the input

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

Let’s take an example

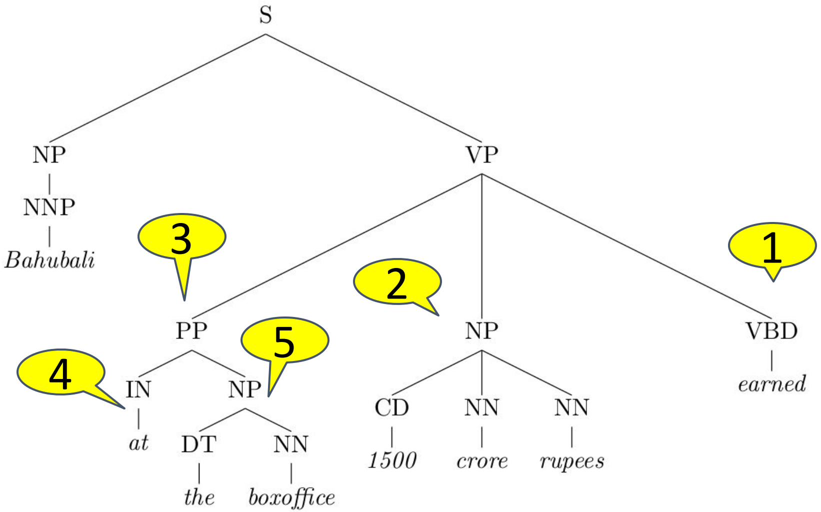
Bahubali earned more than 1500 crore rupee sat the boxoffice

Parse the sentence to understand its syntactic structure



Apply rules to transform the tree

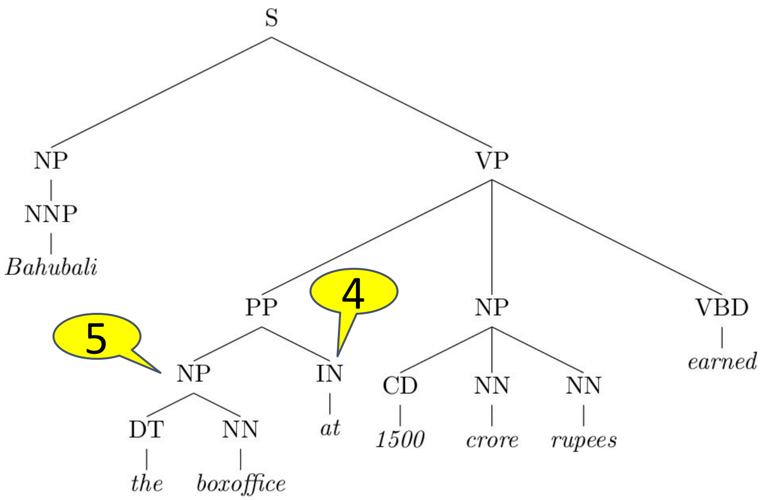
VP ⟶ VBD NP PP ⇒ VP ⟶ PP NP VBD



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

Prepositions in English become postpositions in Hindi

PP ⟶ IN NP ⇒ PP ⟶ NP IN



The new input to the machine translation system is

Bahubali the boxoffice at 1500 crore rupees earned

Now we can translate with little reordering

बाहुबली ने बॉक्सओफिस पर 1500 करोड रुपए कमाए

These rules can be written manually or learnt from parse trees

Better methods exist for generating the correct word order

Incorporate learning of reordering is built into the SMT system

Hierarchical PBSMT ⇒ Provision in the phrase table for limited & simple reordering rules

Syntax-based SMT ⇒ Another SMT paradigm, where the system learns mappings of "treelets" instead of mappings of phrases

We have looked at a basic phrase-based SMT system

This system can learn word and phrase translations from parallel corpora

But many important linguistic phenomena need to be handled

- Divergent Word Order
- **Rich morphology**
- Named Entities and Out-of-Vocabulary words

Language is very productive, you can combine words to generate new words

Inflectional forms of the Marathi word घर

घर	house
घरात	in the house
घरावरती	on the house
घराखाली	below the house
घरामध्ये	in the house
घरामागे	behind the house
घराचा	of the house
घरामागचा	that which is behind the house
घरासमोर	in front of the house
घरासमोरचा	that which is in front of the house
घरांसमोर	in front of the houses

Hindi words with the suffix वाद

साम्यवाद	communism
समाजवाद	socialism
पूंजीवाद	capitalism
जातीवाद	casteism
साम्राज्यवाद	imperialism

The corpus should contains all variants to learn translations

This is infeasible!

Language is very productive, you can combine words to generate new words

Inflectional forms of the Marathi word घर

घर	house
घर ◌ा त	in the house
घर ◌ा वरती	on the house
घर ◌ा खाली	below the house
घर ◌ा मध्ये	in the house
घर ◌ा मागे	behind the house
घर ◌ा चा	of the house
घर ◌ा माग चा	that which is behind the house
घर ◌ा समोर	in front of the house
घर ◌ा समोर चा	that which is in front of the house
घर ◌ा ◌ं	in front of the houses

Hindi words with the suffix वाद

साम्य वाद	communism
समाज वाद	socialism
पूंजी वाद	capitalism
जाती वाद	casteism
साम्राज्य वाद	imperialism

- *Break the words into its component morphemes*
- *Learn translations for the morphemes*
- *Far more likely to find morphemes in the corpus*

We have looked at a basic phrase-based SMT system

This system can learn word and phrase translations from parallel corpora

But many important linguistic phenomena need to be handled

- Divergent Word Order
- Rich morphology
- **Named Entities and Out-of-Vocabulary words**

Some words not seen during train will be seen at test time

These are out-of-vocabulary (OOV) words

***Names** are one of the most important category of OOVs*

⇒ *There will always be names not seen during training*

How do we translate names like Sachin Tendulkar to Hindi?
What we want to do is map the Roman characters to Devanagari to they sound the same when read □ सचिन तेंदुलकर
□ *We call this process **'transliteration'***

How do we transliterate?

Convert a sequence of characters in one script to another script

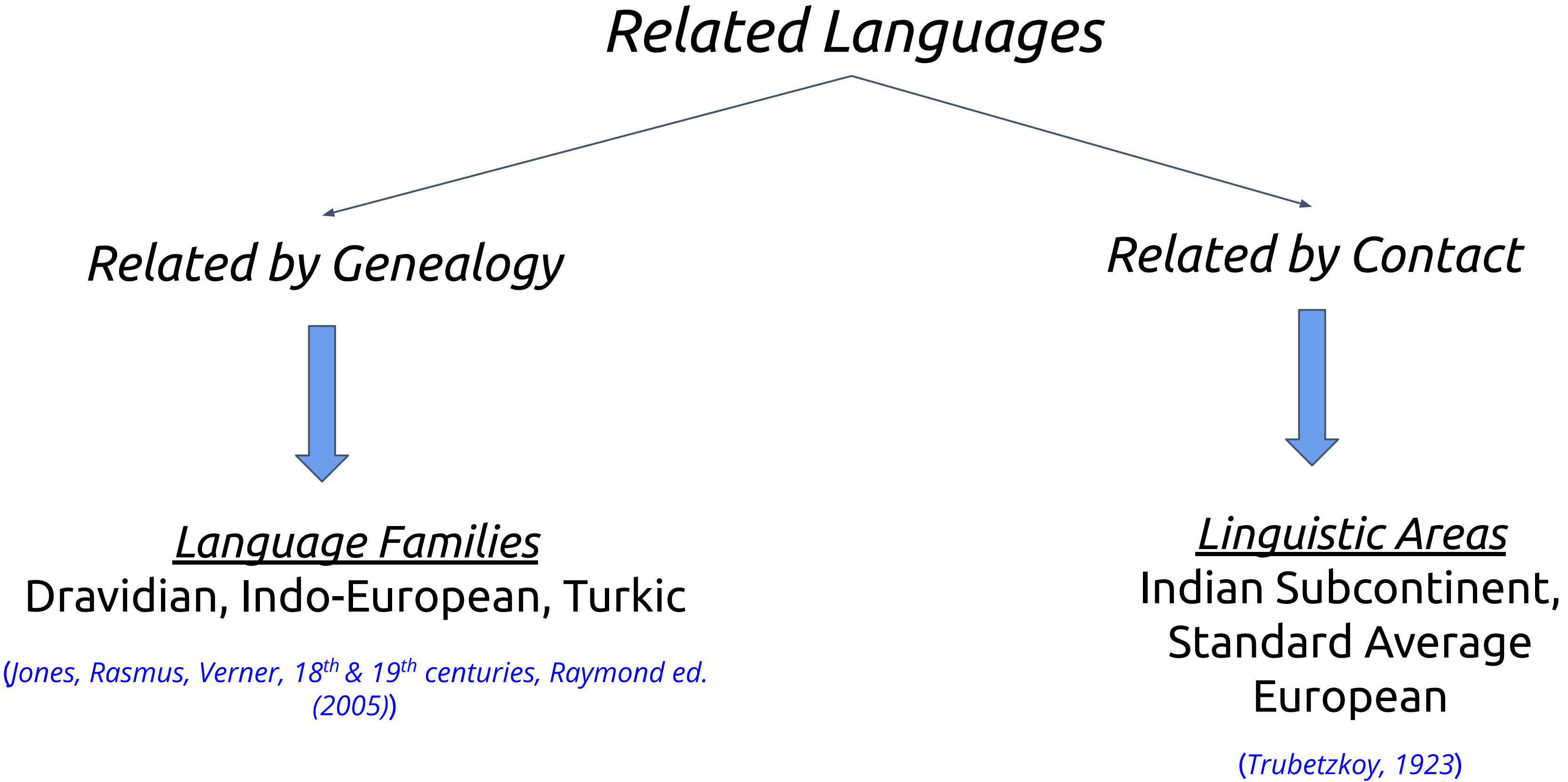
s a c h i n □सच ि न

Isn’t that a translation problem □ at the character level?

Albeit a simpler one,

- *Smaller vocabulary*
- *No reordering*
- *Shorter segments*

Translation between Related Languages



Related Languages

Related by Genealogy

Related by Contact

Language Families
Dravidian, Indo-European, Turkic

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

Linguistic Areas
Indian Subcontinent, Standard Average European

(*Trubetzkoy, 1923*)

Related languages may not belong to the same language family!

Key Similarities between related languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

Marathi

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

Marathi segmented

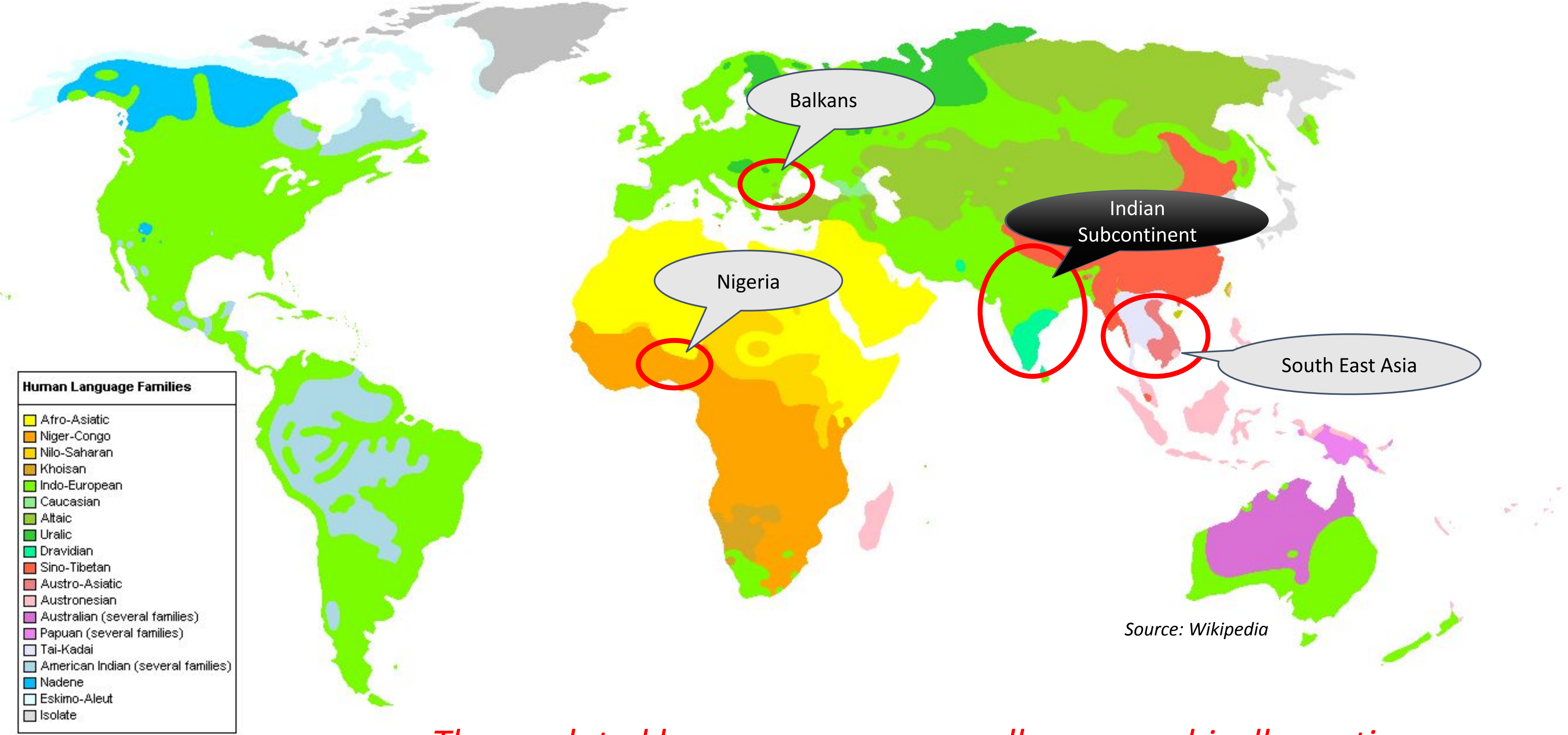
भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

Hindi

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order



Balkans

Indian Subcontinent

Nigeria

South East Asia

Source: Wikipedia

These related languages are generally geographically contiguous

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

Most translation requirements also involves related languages

Between related languages

Hindi-Malayalam
Marathi-Bengali
Czech-Slovak

Related languages ⇐⇒ Link languages

Kannada,Gujarati ⇒ English
English ⇒ Tamil,Telugu

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

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

Lexically Similar Languages

*(Many words having similar **form** and **meaning**)*

- *Cognates*

a common etymological origin

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

- *Loan Words*

borrowed without translation

matsya (sa)	*matsyalu (te)*	*fish*
pazha.m (ta)	*phala (hi)*	*fruit*

- *Named Entities*

do not change across languages

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

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

dAla me.n kuCha kAlA honA	*(hi)*	*Something fishy*
dALa mA kAIka kALu hovu	(gu)	

Translation at subword level which exploits lexical similarity

What is a good unit of representation?

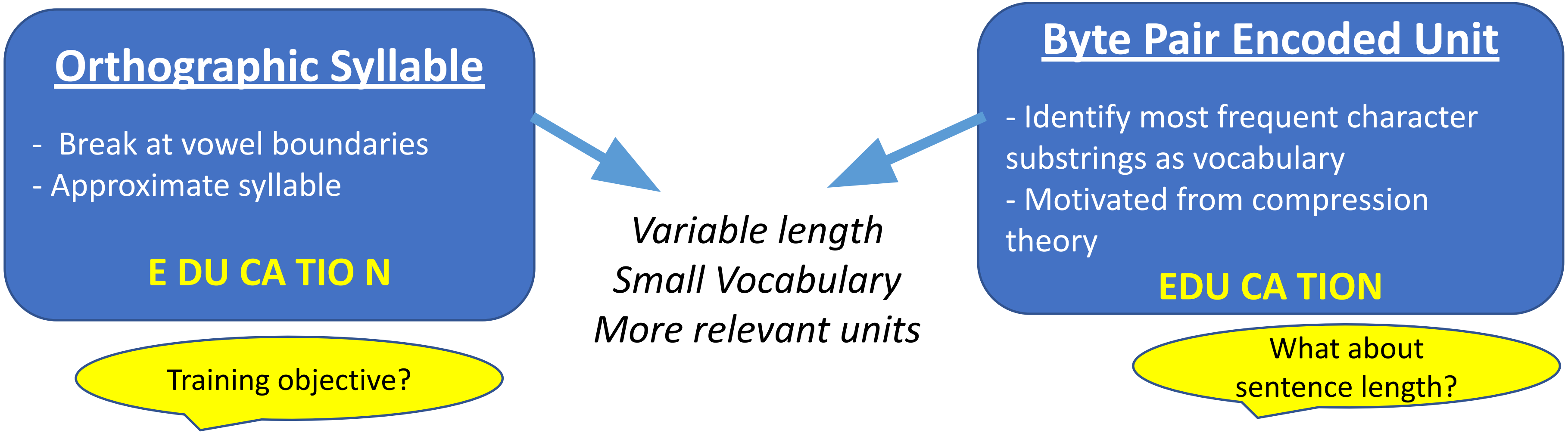
Let's take the word EDUCATION as an example

<u>Character</u>: E D U C A T I O N

ambiguity in character mappings

<u>Character n-gram</u>: ED UC AT IO N

Vocabulary size explodes for n>2



<u>Orthographic Syllable</u>

- Break at vowel boundaries
- Approximate syllable

E DU CA TIO N

<u>Byte Pair Encoded Unit</u>

- Identify most frequent character substrings as vocabulary
- Motivated from compression theory

EDU CA TION

Variable length

Small Vocabulary

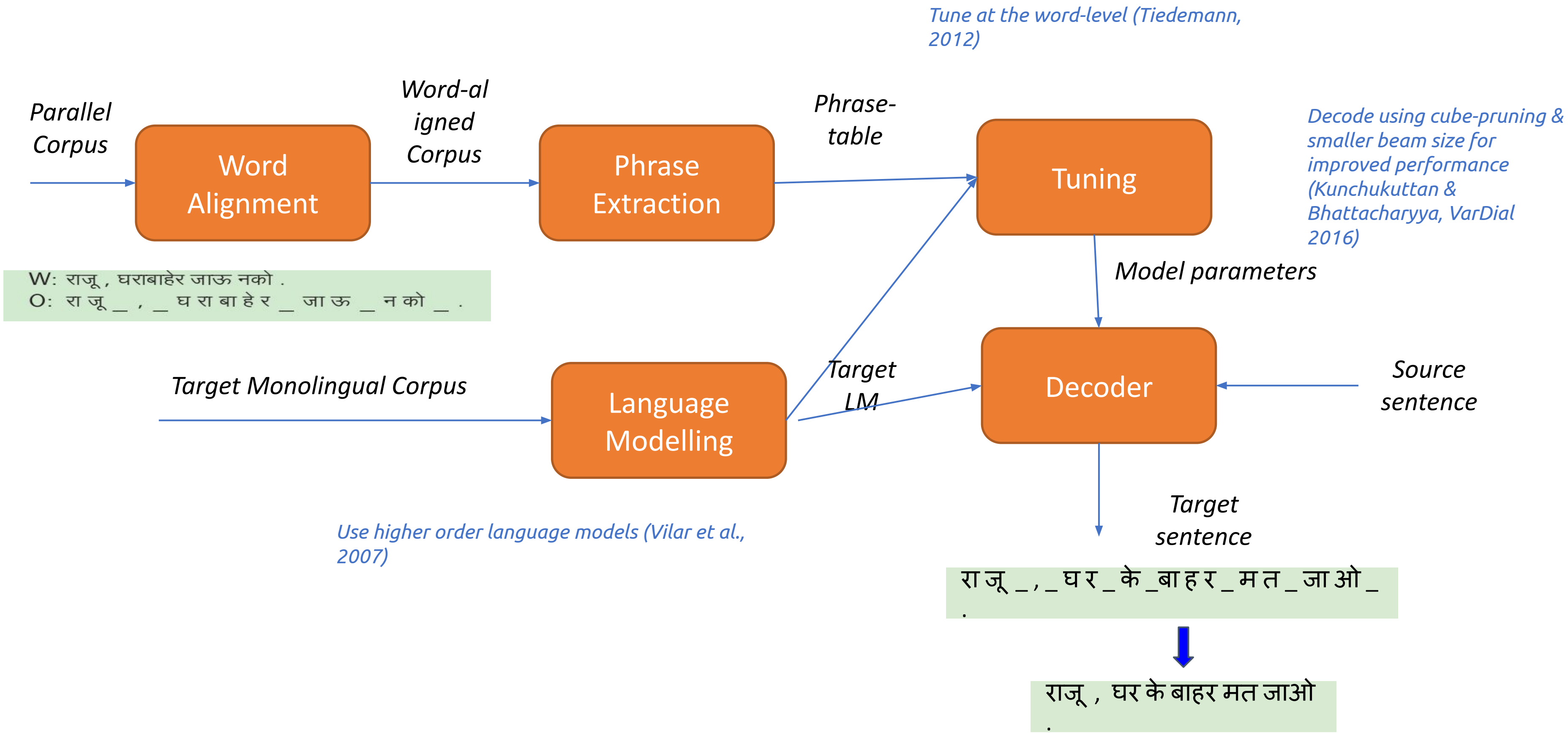
More relevant units

Training objective?

What about sentence length?

<u>Sentence Representation</u> → मु म्ब ई _ म हा रा ष्ट्र _ की _ रा ज धा नी _ है _ ।

Adapting SMT for subword-level translation



Tune at the word-level (Tiedemann, 2012)

Decode using cube-pruning & smaller beam size for improved performance (Kunchukuttan & Bhattacharyya, VarDial 2016)

Use higher order language models (Vilar et al., 2007)

W: राजू , घराबाहेर जाऊ नको .
O: रा जू _ , _ घ रा बा हे र _ जा ऊ _ न को _ .

रा जू _ , _ घ र _ के _बा ह र _ म त _ जा ओ _
.

राजू , घर के बाहर मत जाओ
.