On Teaching and Learning Machine Translation

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay and IIT Patna

ACL 16, Berlin

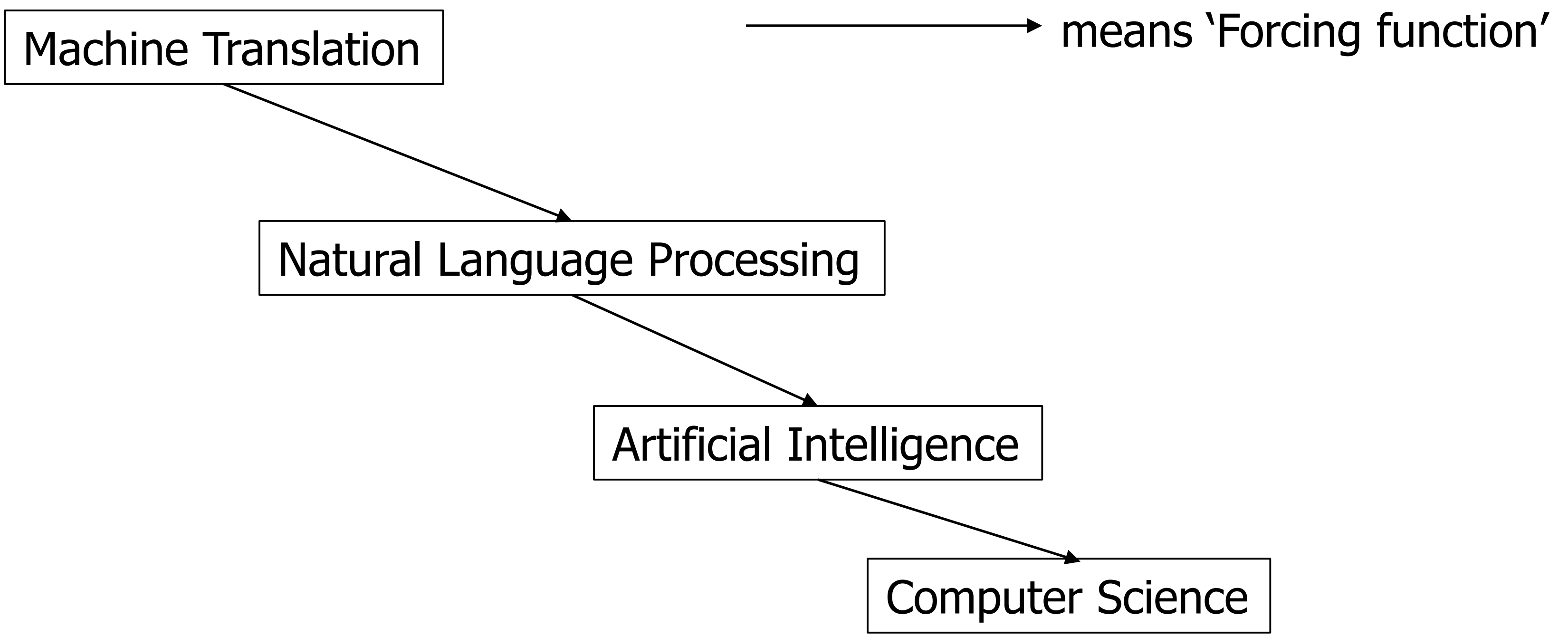
10th August 2016

President’s talk- expectations

- Scintillating talks from past presidents- instructive, humorous, setting agenda

- I thought of dwelling on MT teaching and learning
  - Complements MT research

Why in particular Machine Translation? *Forcing Function*



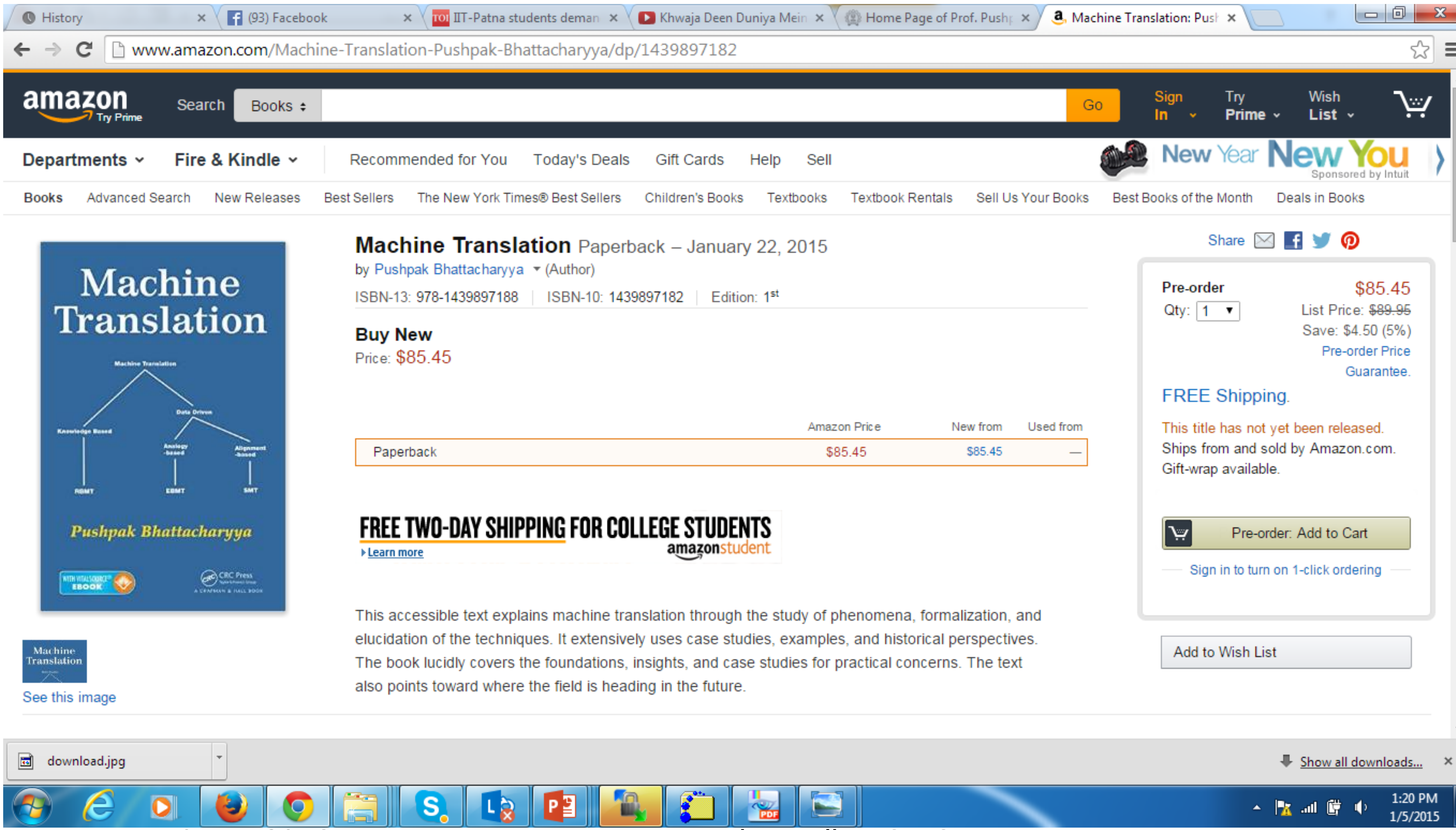
Making the makers of machine translation

- Practitioners of Machine Translation make and study machines that translate between languages

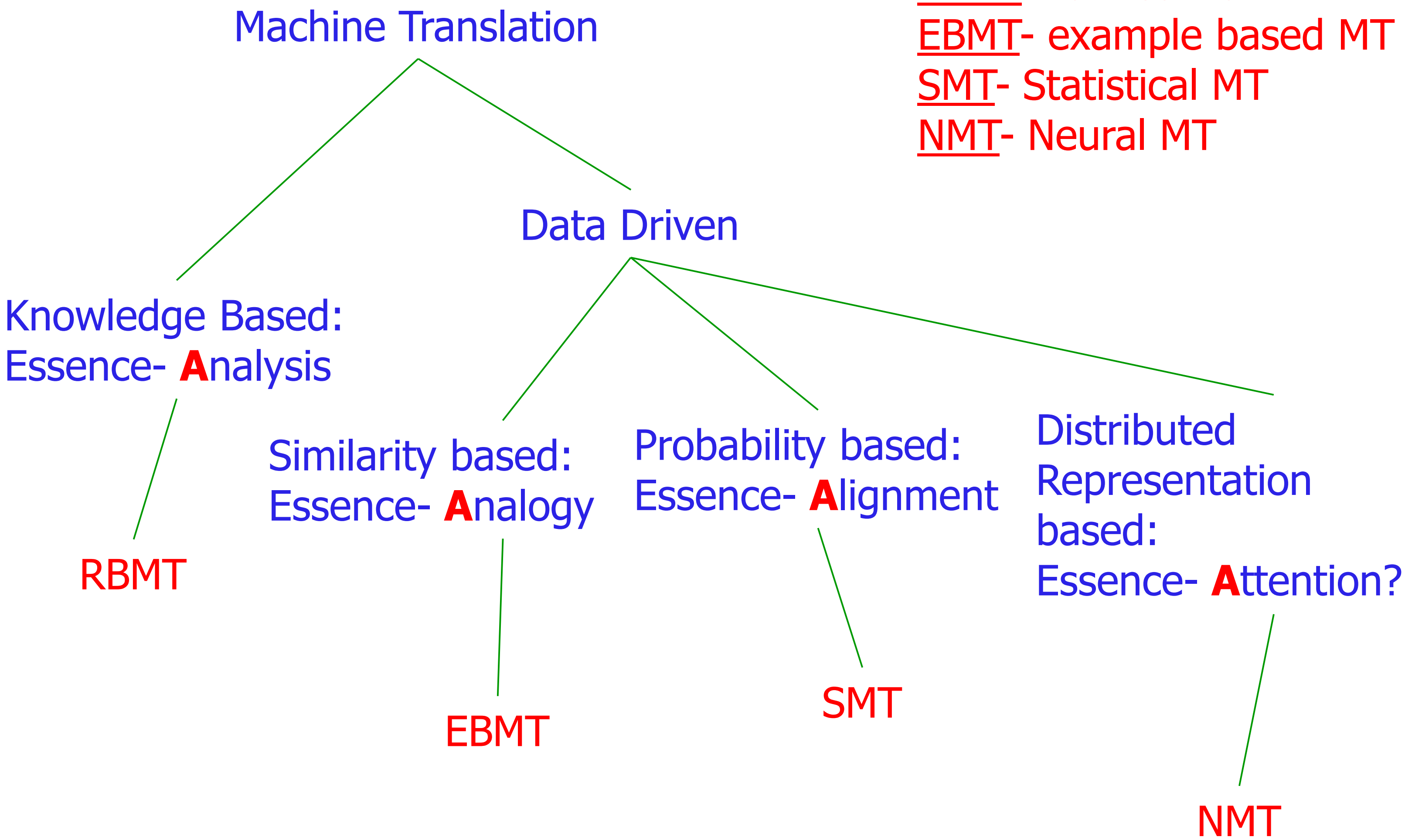
- How to teach MT to the would be practitioners?

Book resulting from 25 years of teaching and researching in NLP and MT

http://www.amazon.com/Machine-Translation-Pushpak-Bhattacharyya/dp/1439897182



RBMT- rule based MT
EBMT- example based MT
SMT- Statistical MT
NMT- Neural MT



Questions

- What unifies these approaches of RBMT, EBMT, SMT and now NMT?
- Why did these approaches meet with varied amounts of success?
- What are the essential elements of these approaches?

Very specific 2 questions (1/2)

- Within RBMT, why has transfer based MT been more successful than interlingua based MT?
  - Case of SYSTRAN (transfer based)
  - Our experience of PAN Indian Indian Language to Indian Language MT (transfer based)

Very specific 2 questions (2/2)

- Though both are data driven, why did EBMT's journey languish midway, while SMT kept progressing?
  - Questions like the above are important for teaching, learning and doing research in MT

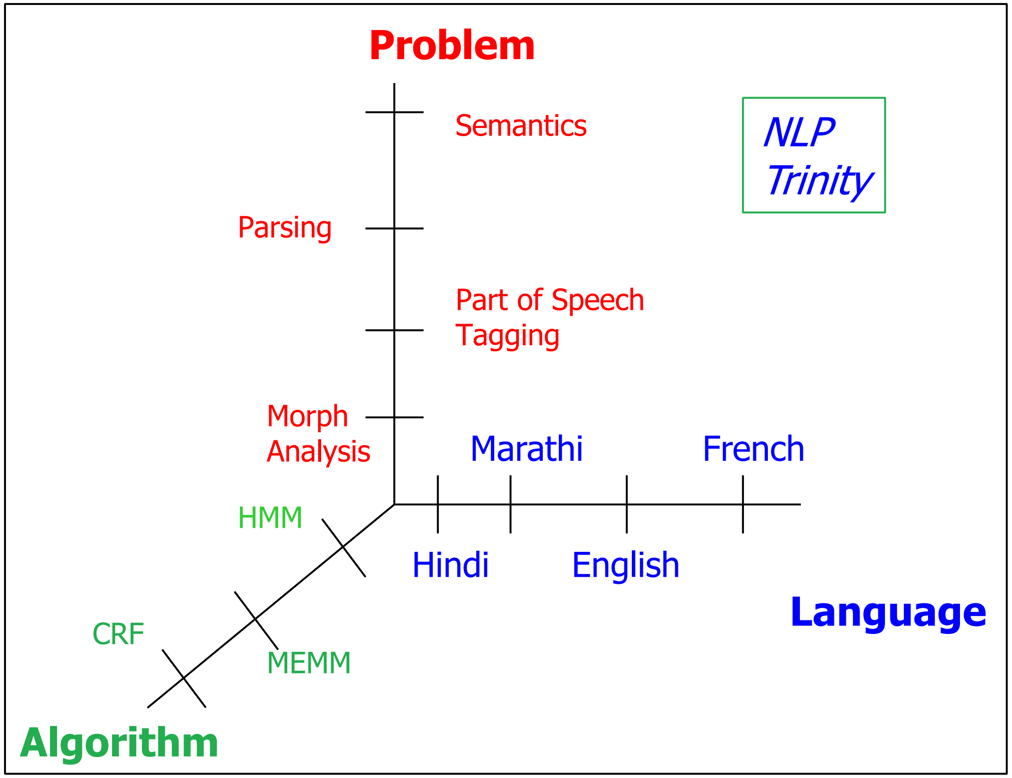
Plausible answers: based on the nature of MT approaches

- ONE thing RBMT should do well:
  - Answer- **A**nalysis
- One thing SMT should do well:
  - Answer- **A**lignment
- One thing EBMT do well:
  - Answer- **A**nalogy
- One thing NMT should do well:
  - Answer- **A**ttention?

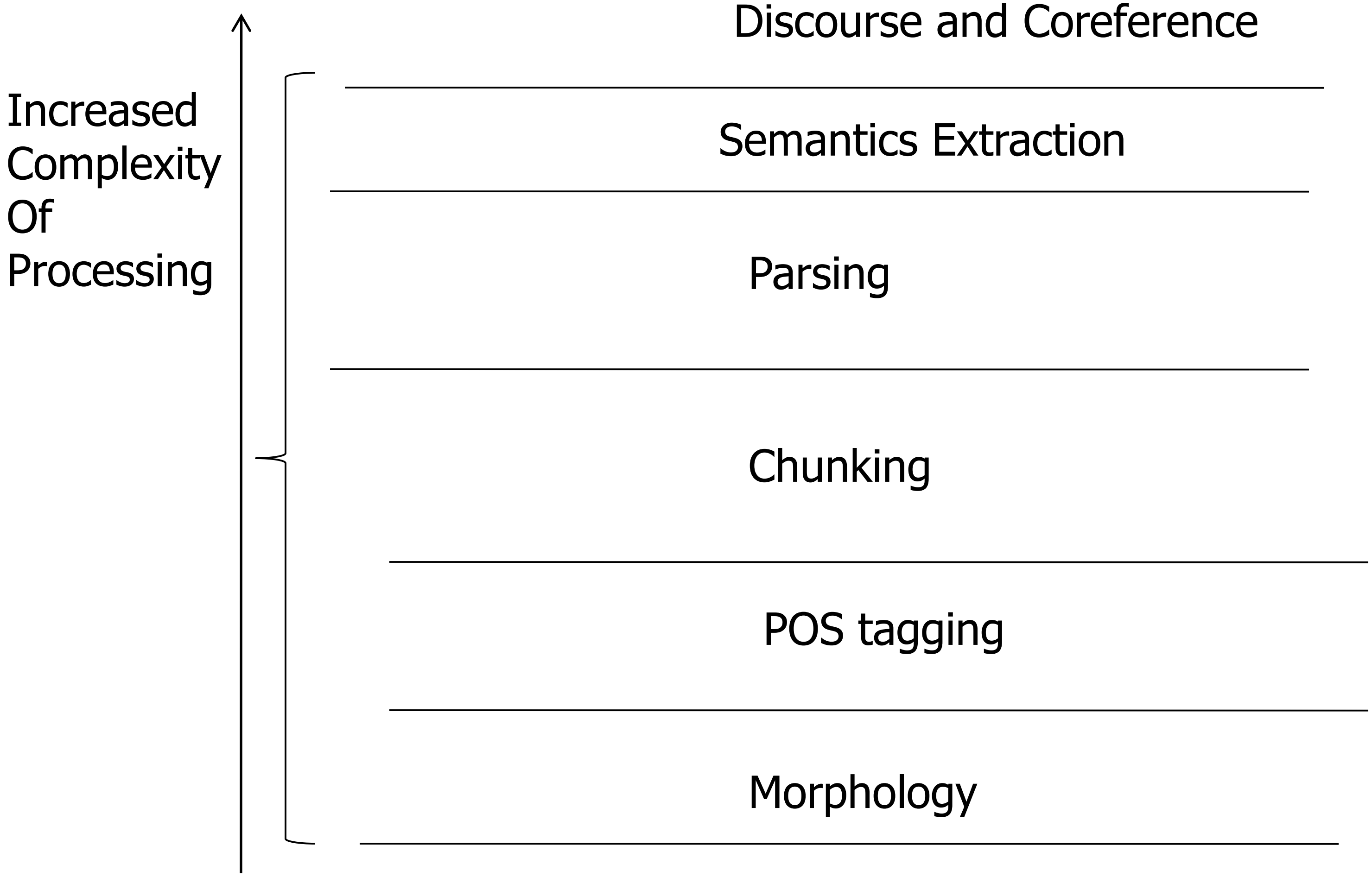
Main message of this presentation

- To teach, learn, and do research in MT, one should **FIRST** note and understand the role of
  - **A**nalysis in RBMT
  - **A**lignment in SMT
  - **A**nalogy in EBMT
  - **A**ttention in NMT?

MT is a part of NLP which is a Trinity

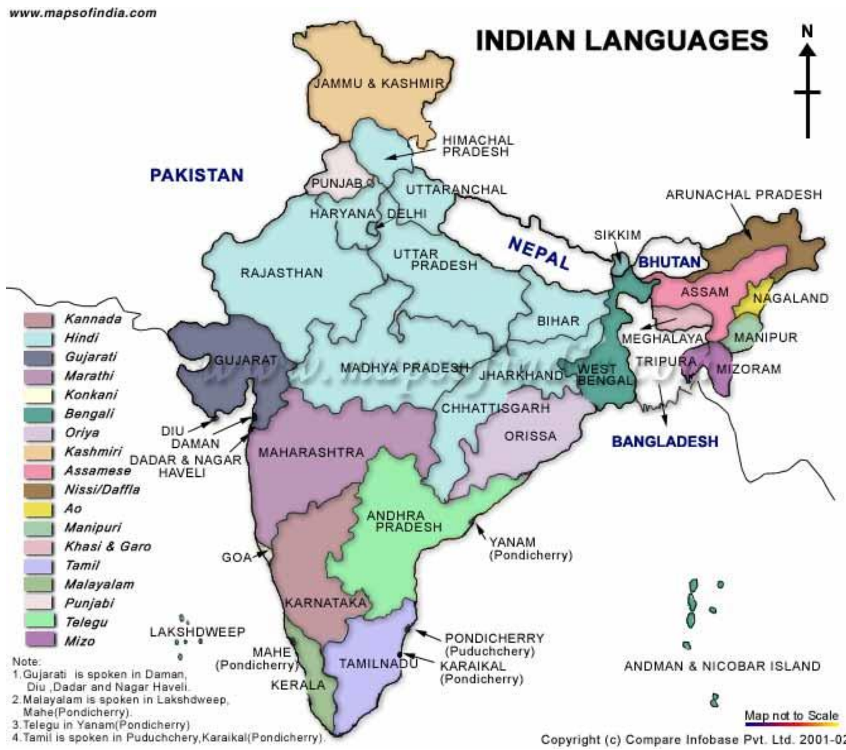


NLP Layer/Pipeline



Multilinguality- a way of life in India

- People speak, read, write 3 languages routinely
- Code mixing and language switch are a way of life
- Major streams:
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Large speaker population
  - Hindi and Urdu: 5th (~500 milion)
  - Bangla: 7th (~300 million)
  - Marathi 14th (~70 million)



Why is MT difficult: Language Divergence

- Remember the trap that MT fell into in the 60s?- MT is just dictionary look up and substitution
- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

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

Different ways of expressing meaning

Manipuri:

kampor	*asi*	*mon*	*mon*	*laui*
blanket	*this*	*soft*	*soft*	*is*

English:

This blanket is very soft

Hindi:

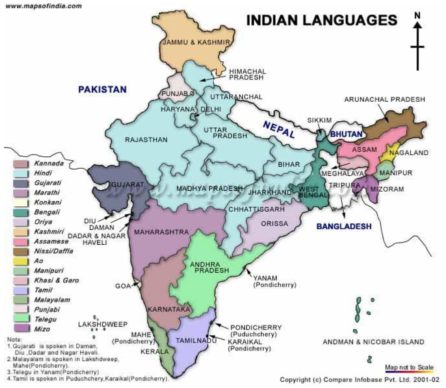
yaha kambal bahut naram hai

Bangla:

ei kambal ti khub naram

Marathi:

haa kambal khup naram aahe



Languages differ in expressing thoughts: Finnish agglutination

Finnish*:* “istahtaisinkohan”

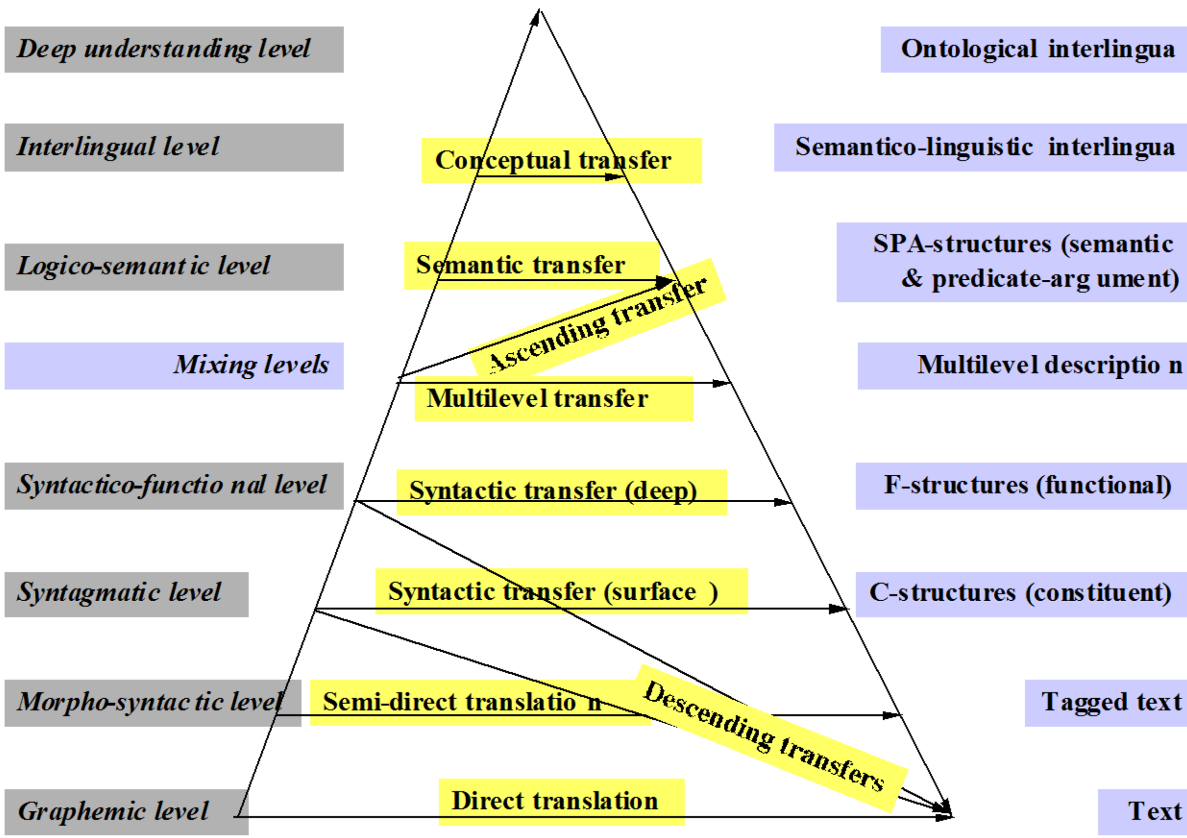
English: "I wonder if I should sit down for a while“

Analysis:

- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix
- n + 1st person singular suffix
- ko + question particle
- han a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)

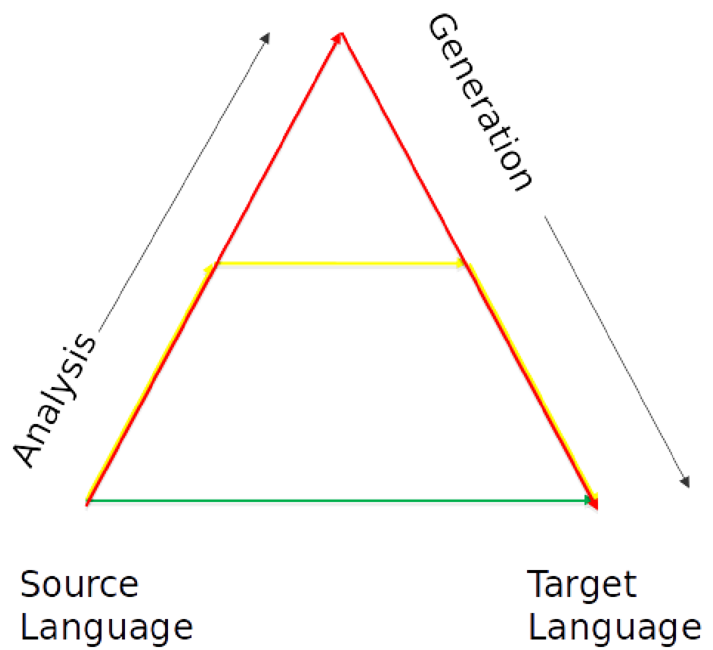
Kinds of MT Systems

(point of entry from source to the target text)



(Vauquois. 1968)

Simplified Vauquois



Interlingua Based Translation

Transfer Based Translation

Direct Translation

RBMT-EBMT-SMT spectrum: knowledge (rules) intensive to data (learning) intensive

Illustration of difference of RBMT, SMT, EMT

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

The tricky case of ‘have’ translation

English

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

Marathi

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

RBMT

If

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

Then

"have" should translate to "kade … aahe"

If

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

Then

"have" should translate to "laa … aahe"

If

syntactic subject is **inanimate**

Then

"have" should translate to "madhye … aahe"

EBMT

X ***have*** *Y* →

X_kade Y aahe /

X_laa Y aahe /

X_madhye Y aahe

SMT

- *has a house ←→ kade ek ghar aahe*

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

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

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

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

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

  hospital <cm> has

SMT: new sentence

"This hospital has 100 beds"

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

DECODING !!!

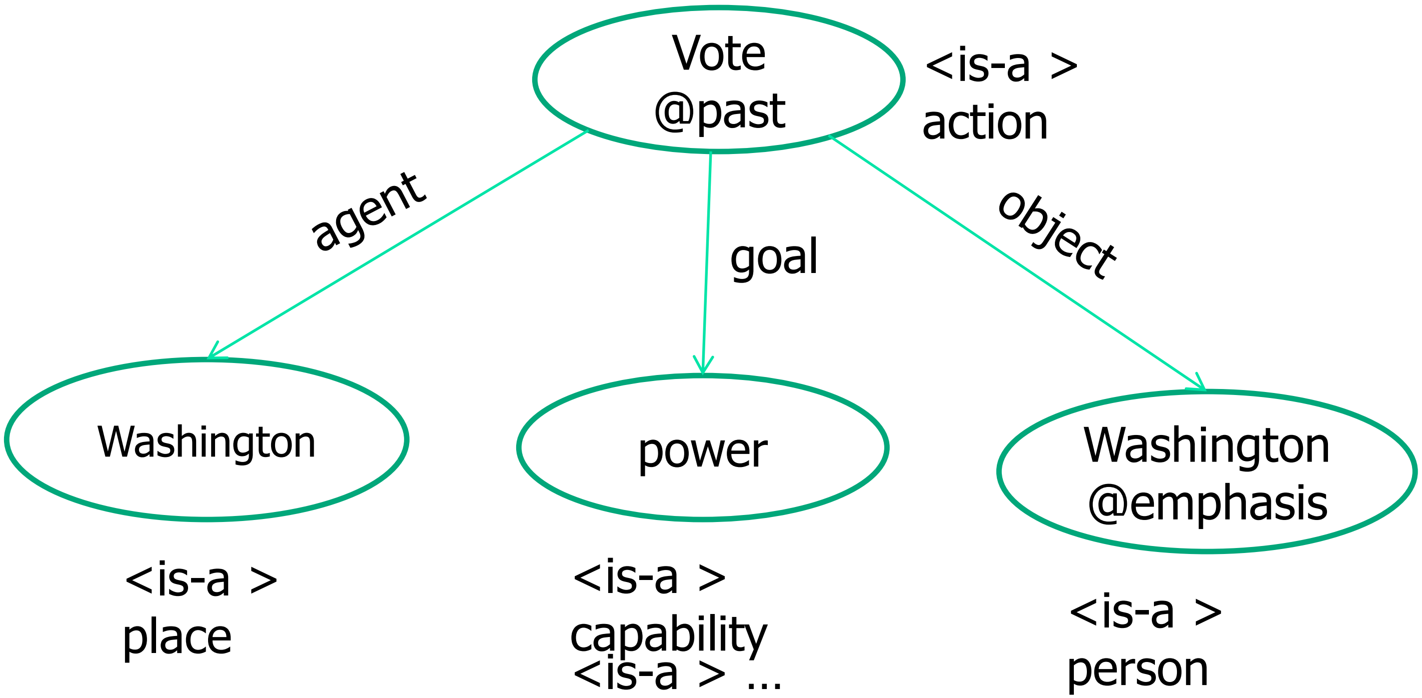
Essentials of RBMT

Interlingua based MT: Universality hypothesis

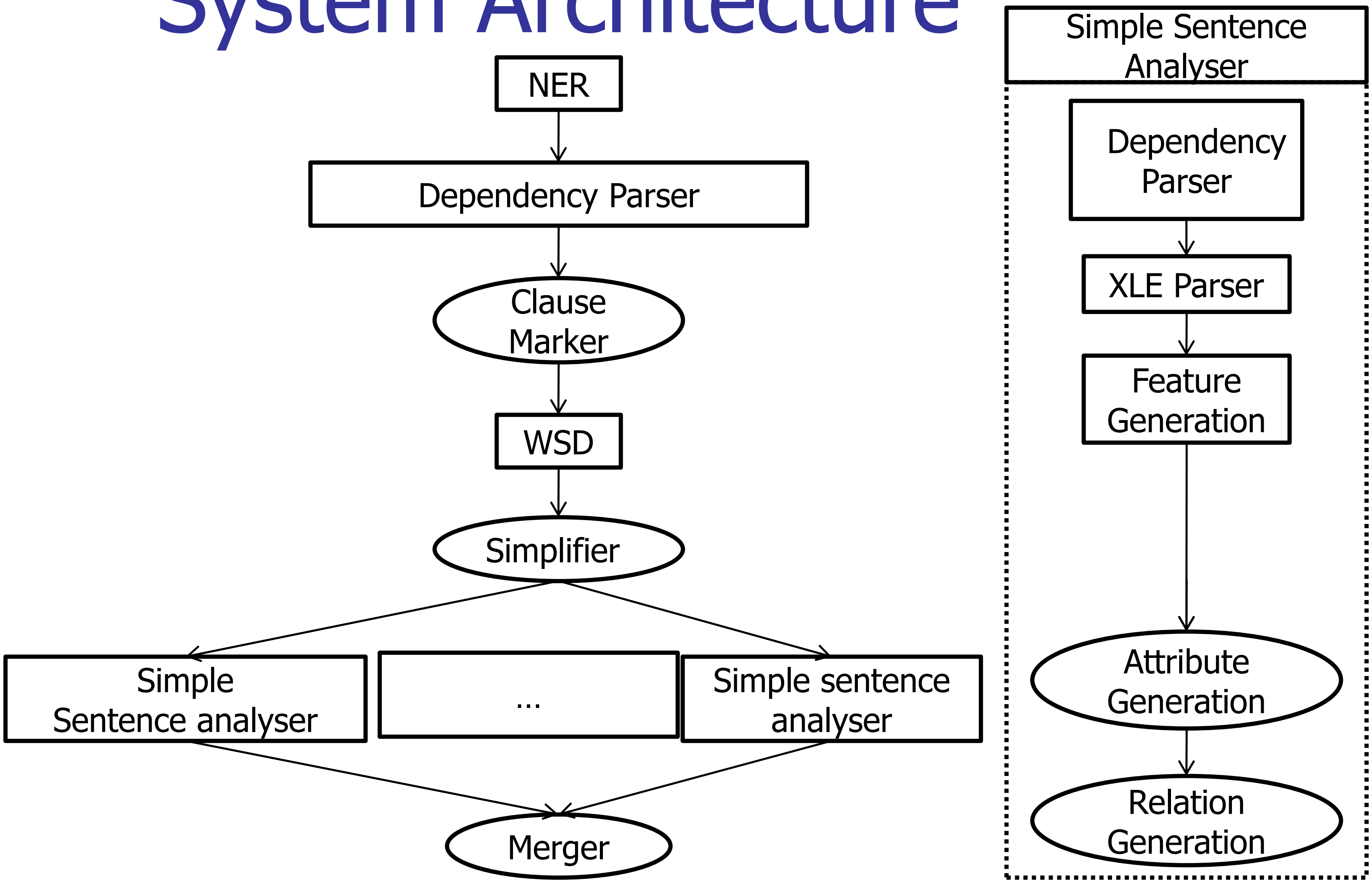
Universality hypothesis: At the level of "deep meaning", all texts are the "same", whatever the language.

Interlingua based MT: semantic parsing- complete disambiguation

Washington, Washington voted to power, heralding new hopes and aspirations.

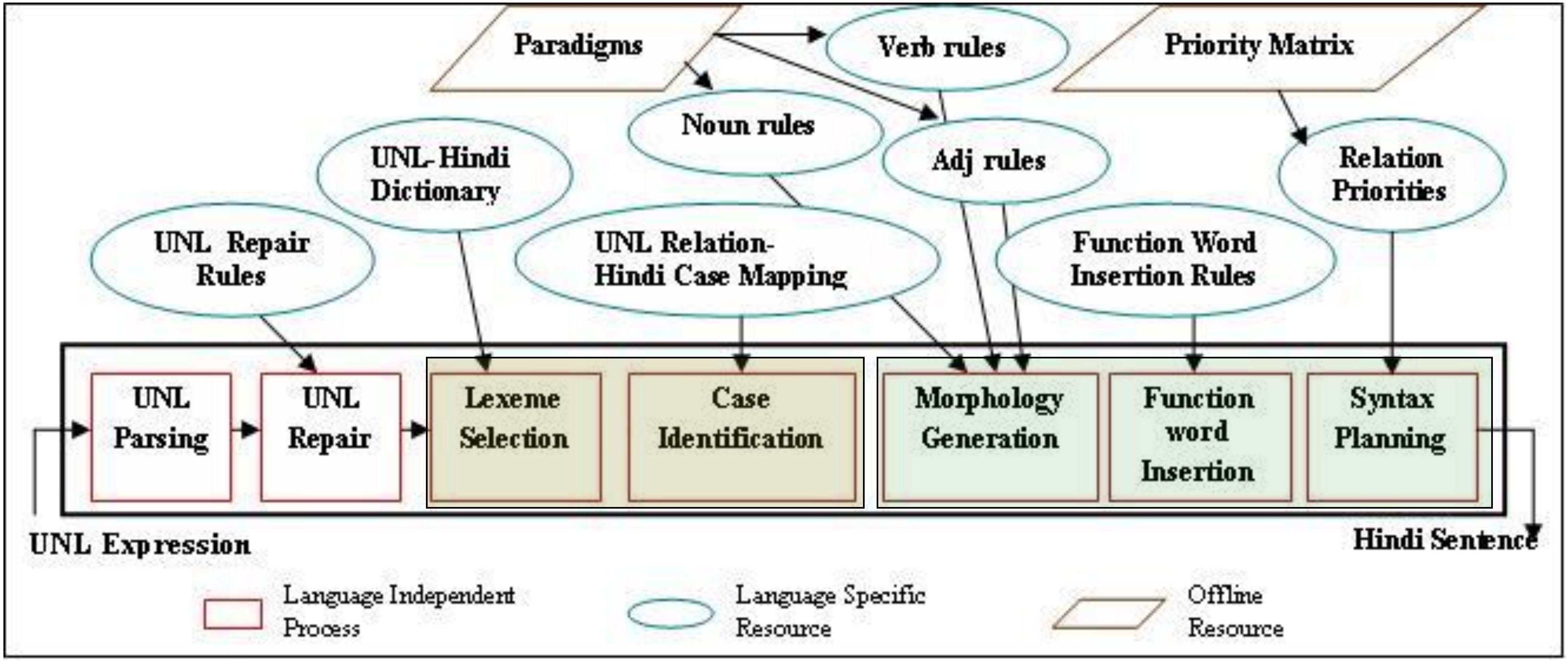


System Architecture



Generation Architecture (with example from Universal Networking Language Project)

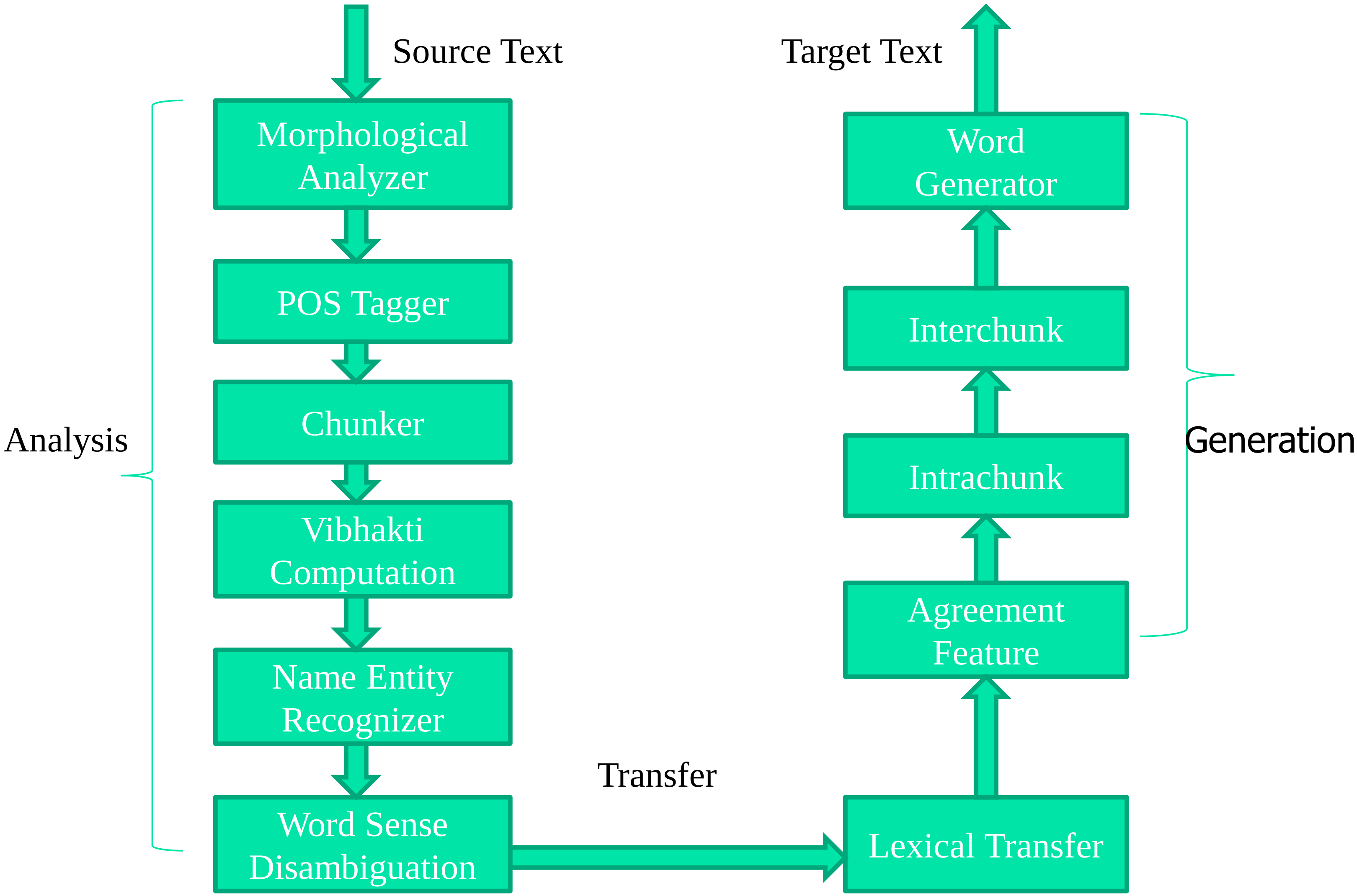
Deconversion = Transfer + Generation



Transfer based MT: Indian Language to Indian Language Machine Translation (ILILMT)

- Bidirectional Machine Translation System
- Developed for nine Indian language pairs
- Approach:
  - Transfer based
  - Modules developed using both rule based and statistical approach

Architecture of ILILMT System



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

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

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

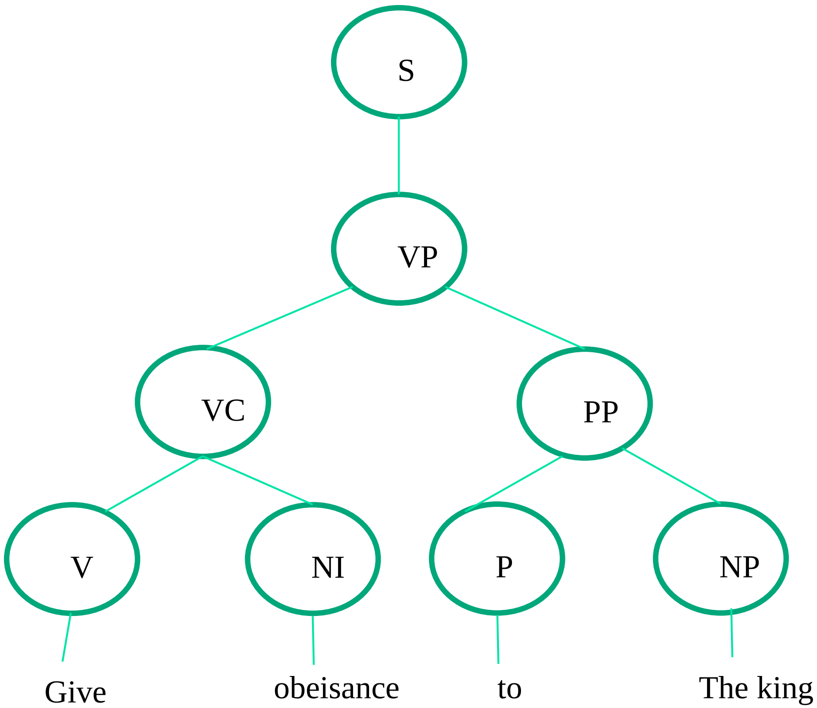
- **அரசரை வணங்கு** (Tamil; Dravidian)
- aracarai vanaNku
- king_to obeisance_do

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

transfer amongst different language families

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

English parse tree



Transfer rules:

- VC-PP inversion (all languages)

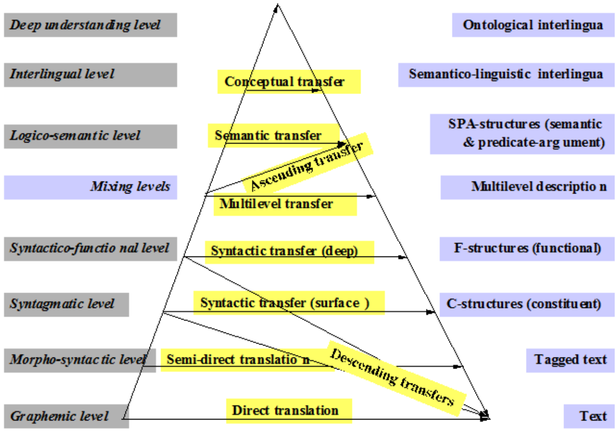
VC

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

PP

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

Essential element of Statistical Machine Translation



Linguistics 101: Czeck-English data

- [nesu] “I carry”
- [ponese] “He will carry”
- [nese] “He carries”
- [nesou] “They carry”
- [yedu] “I drive”
- [plavou] “They swim”

To translate …

- I will carry.
- They drive.
- He swims.
- They will drive.

Thought process of students

- Naturally begin to align
- Segment strings
- Establish correspondences
- Select and eliminate possibilities

Foundation

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

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

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

The all important **word alignment**

- The edifice on which the structure of SMT is built (Brown et. Al., 1990, 1993; Och and Ney, 1993)
- Word alignment → Phrase alignment (Koehn et al, 2003)
- Word alignment → Tree Alignment (Chiang 2005, 200t; Koehn 2010)
- Alignment at the heart of Factor based SMT too (Koehn and Hoang 2007)

Word alignment is the crux of the matter: how to convey the essential idea?

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

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

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

"counts"

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

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

Revised probabilities table

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

“revised counts”

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

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

Re-Revised probabilities table

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

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

Derivation: Key Notations

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

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

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

.....

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

.....

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

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

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

Modeling: Hidden variables and parameters

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

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

Likelihoods

Data Likelihood *L(D; Θ)* :

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

Data Log-Likelihood LL(D; Θ) :

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

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

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

Constraint and Lagrangian

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

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

Differentiating wrt P_{ij}

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

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

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

Final E and M steps

M-step

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

E-step

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

Essential element of Example based Machine Translation

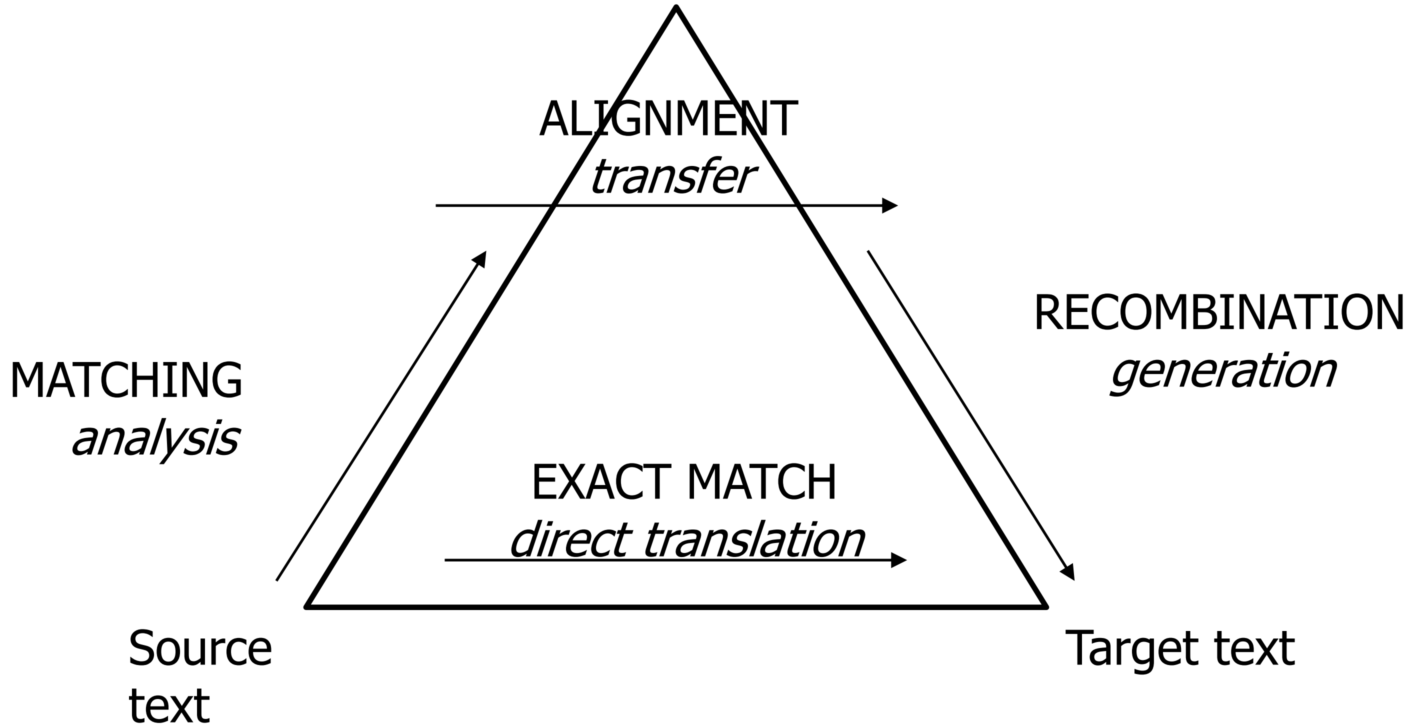
Nagao’s seminal paper 1984 (1/2)

“Man does not translate a simple sentence by doing ***deep linguistic analysis****,* rather,

man does the translation, first, by properly decomposing an input sentence into certain ***fragmental phrases*** (very often, into case frame units), and then (p.t.o)

Nagao’s seminal paper 1984 (2/2)

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



The “Vauquois pyramid” adapted for EBMT

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

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

- Large lexical knowledge networks that provide resources for measuring similarity.

Different ways of measuring text similarity

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

Bag of words

- $B(S_1)$ and $B(S_2)$ which are bags of words, i.e., sets of word in the two sentences

$$Dice(B(S_1), B(S_2)) = \frac{|B(S_1) \cap B(S_2)|}{|B(S_1)| + |B(S_2)|}$$

$$Jackard(B(S_1), B(S_2)) = \frac{|B(S_1) \cap B(S_2)|}{|B(S_1) \cup B(S_2)|}$$

N-gram based matching: BLEU score

Recall -> Brevity Penalty

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

Precision -> Modified n-gram precision

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

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

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

Vector based

- Vectors of 1/0, term presence or absence, or
- Vectors of term frequencies (TF), or
- Vectors of TF*IDF, where IDF is the inverse document frequency
- Finally use cosine similarity

$$\cos ine(V(S_1), V(S_2)) = \frac{V(S_1).V(S_2)}{|V(S_1)|.|V(S_2)|}$$

$$idf(w) = \log\left(\frac{N}{|S;\ w \in S|}\right)$$

where, N is the total number of sentences. the denominator inside 'log' gives the total number of sentences containing the word 'w'.

Tree based: Constituency

- N_1: number of nodes in S_1's constituency tree
- N_2: number of nodes in S_2's constituency tree.
- M: number of nodes matched in a particular order of traversal (pre-, in- or post- etc.)

$S(S_1, S_2)$: similarity of S_1 and S_2 =

$$\frac{M}{\max(N_1, N_2)}$$

Tree based: Dependency

- D_1: number of dependency relations in S_1
- D_2: number of dependency relations in S_2
- w_r: weight of matching the relation part (same for all relations)
- w_{arg1}: weight of matching the first argument (same for all relations)
- w_{arg2}: weight of matching the 2nd argument (same for all relations)
- *Rs*: relations in two trees
- *As*: first arguments in the relations
- *Bs*: second arguments in the relations

$$S(S_1, S_2) = \frac{\sum_{i=1}^{|D_1|} \sum_{j=1}^{|D_2|} [w_r \delta(R_i^1, R_j^2) + w_{\arg 1} \delta(A_1^i, A_1^i) + w_{\arg 2} \delta(B_1^i, B_2^j)]}{\max(D_1, D_2)}$$

Feature based (very rich)

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

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

Deep semantics based

- Word meaning
- Semantic relations
- Speech acts (tense, number and other pragmatic attributes like focus, emphasis etc.)
- Use of text entailment

EBMT's 'decoding': RECOMBINATION

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

Example of re-instantiation

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

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

Re-instantiation: adjustments (boundary friction problem)

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

Re-instantiation leading to translation

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

Summary (1/2)

- To teach MT: emphasise the 3 'A's as the "first thing first"
  - RBMT- **A**nalysis (NLP pipeline)
  - SMT- **A**lignment (especially, IBM Model 1)
  - EBMT- **A**nalogy (deep matching)

- Compare and contrast the approaches: the taxonomy diagram

Summary (2/2)

- Analyse critically the historical development: presence and absence of tools and resources, computing power and so on and their effect

- Emphasise and re-emphasise the seminal role of EM based word alignment

- Emphasise and reemphasise the fact that transfer based MT can CHOOSE the level of transfer unlike that of pure interlingua based MT

These observations are based on (1/2)

- 25 years of teaching and researching NLP and MT
- Book writing (2010-2015)
- Multiple projects on MT
  - UNL (1996-2000): **interlingua based**, UNU funded
  - Sampark (Indian language to Indian language MT, 2000-till date): **Transfer based**, Ministry of IT funded "Technology Development in IL"

These observations are based on (2/2)

- Multiple projects on MT
  - Anuvadaksh (English to Indian language MT, 2006- till date) **SMT** based, Ministry of IT funded “Technology Development in IL”
  - MT Projects sponsored by Elsevier (English-Bahasa, English-Marathi): **SMT with Pivot and Factors**
  - MT Projects sponsored by Accenture (Eng-Hin-Mar): **Hybrid**

And frequent observations like this

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

Baseline PBSMT - % BLEU scores (S1)

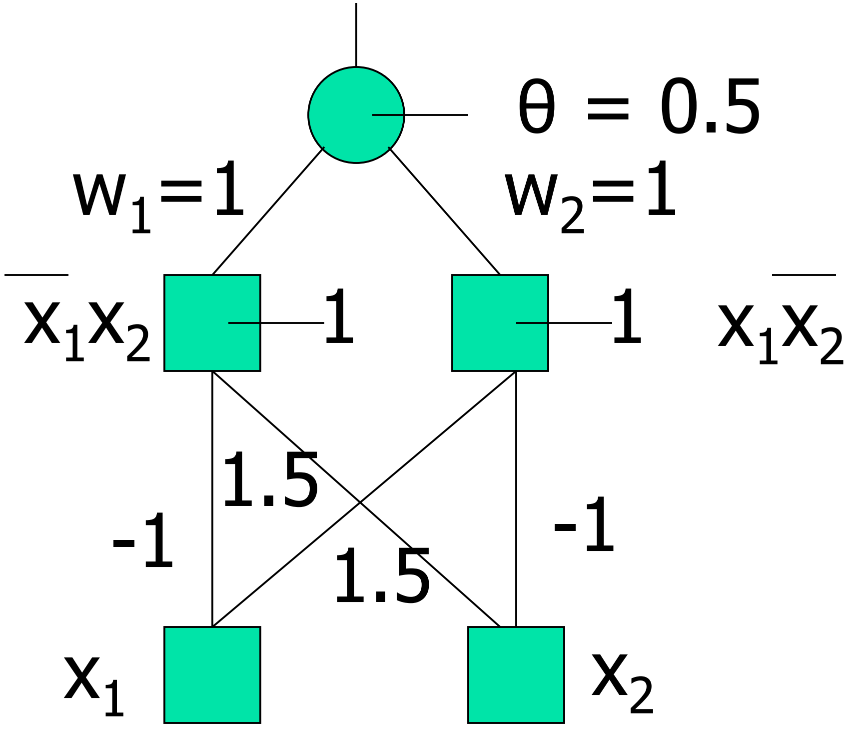
- **Clear partitioning of translation pairs by language family pairs**, based on translation accuracy.
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult

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

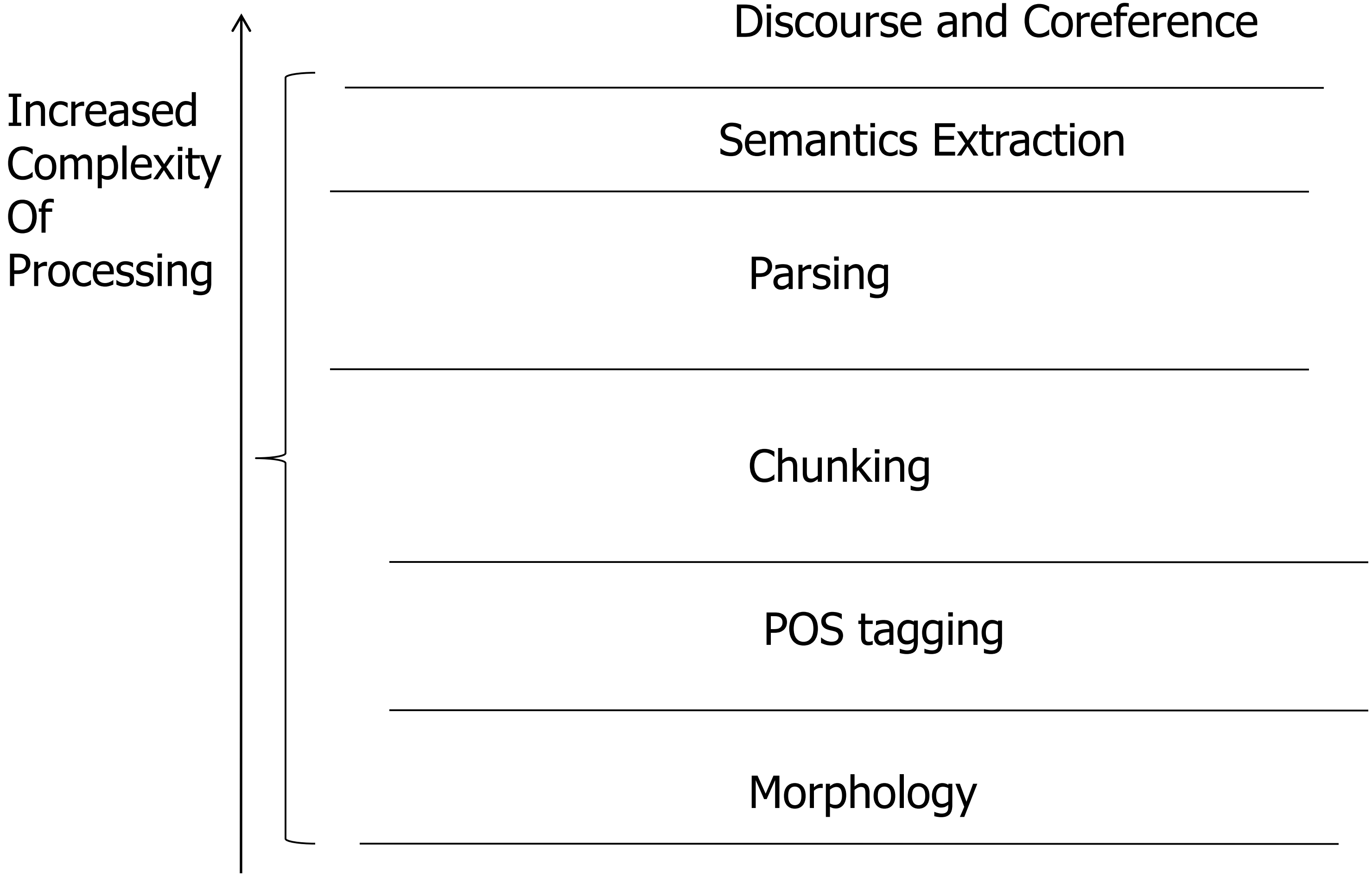
What about neural machine translation?

- NMT a particular instance of solving mapping problems by neural networks
- Remember Feedforward Neural Nets?
- Hidden layer does amazing things

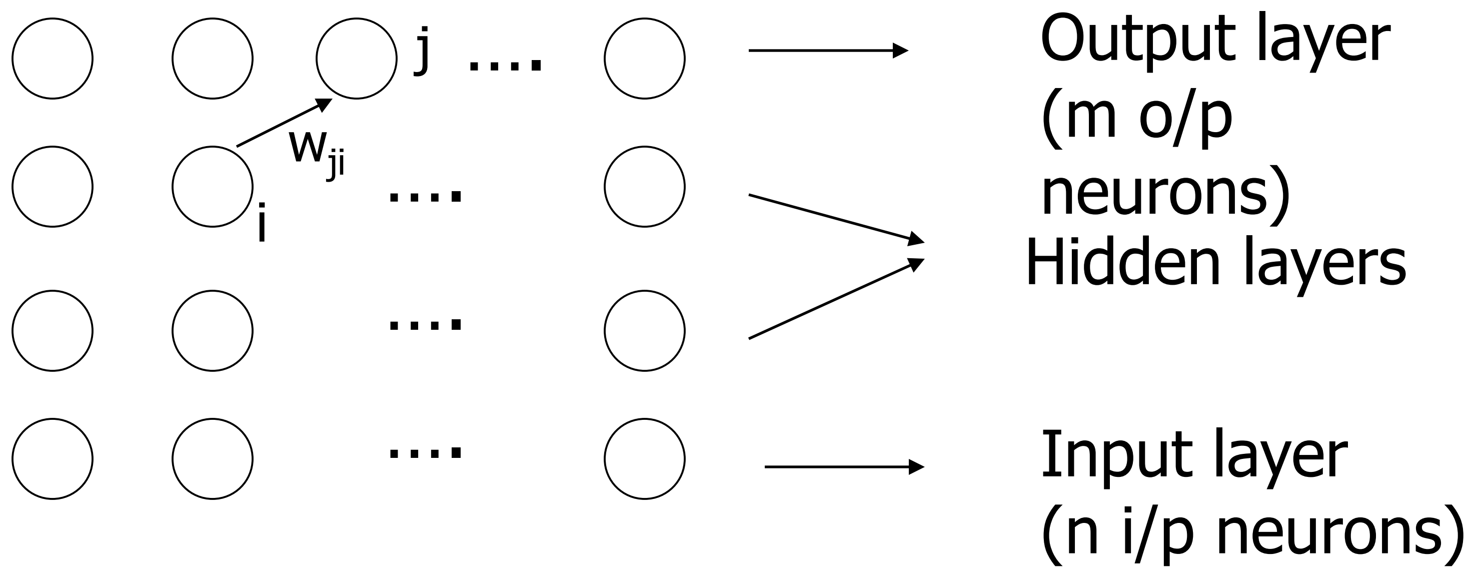
Example – XOR



NLP Layer/Pipeline



Multilayer neural net



- NLP pipeline ←→ NN layers
- Discover bigger structures bottom up, starting from character?
- Words, POS, Parse, Sentence, Discourse?

What about neural machine translation? (contd.)

- NMT a particular instance of solving mapping problems by neural networks
- Spectacular success in speech and vision (as high as 50% reduction in error rate)

What about neural machine translation? (contd.)

- Not such spectacular success in NLP, though
  - Why?
  - Any particular point whose time is yet to come? (remember EBMT?)
  - Any core algorithm waiting to be developed (remember EM based alignment in SMT?)
  - Need to go deeper into **A**ttention? Another 'A'

Thank you

Ongoing work (1/2)

- Cognitive NLP for MT
  - *John finds it difficult to handle the situation*
  - *John is in a soup*
- Automatic correction of machine translation output- post editing
- Injection of morphological information in the training corpus
- Pivot study

NLP@IITB- since 1996 at CSE Dept.

<u>Information Extraction:</u>
- **Part of Speech tagging**
- **Named Entity Recognition**
- **Shallow Parsing**
- **Summarization**

<u>IR:</u>
- **Cross Lingual Search**
- **Multilingual Relevance Feedback**

Language Processing & Understanding

Linguistics is the eye and computation the body

<u>Machine Learning for NLP</u>
- **Semantic Role labeling**
- **Sentiment Analysis**
- **Text Entailment**

Using graphical models, support vector machines, neural networks

<u>Machine Translation:</u>
- **Statistical**
- **Interlingua Based**
- **English→Indian languages**
- **Indian languages→Indian languages**

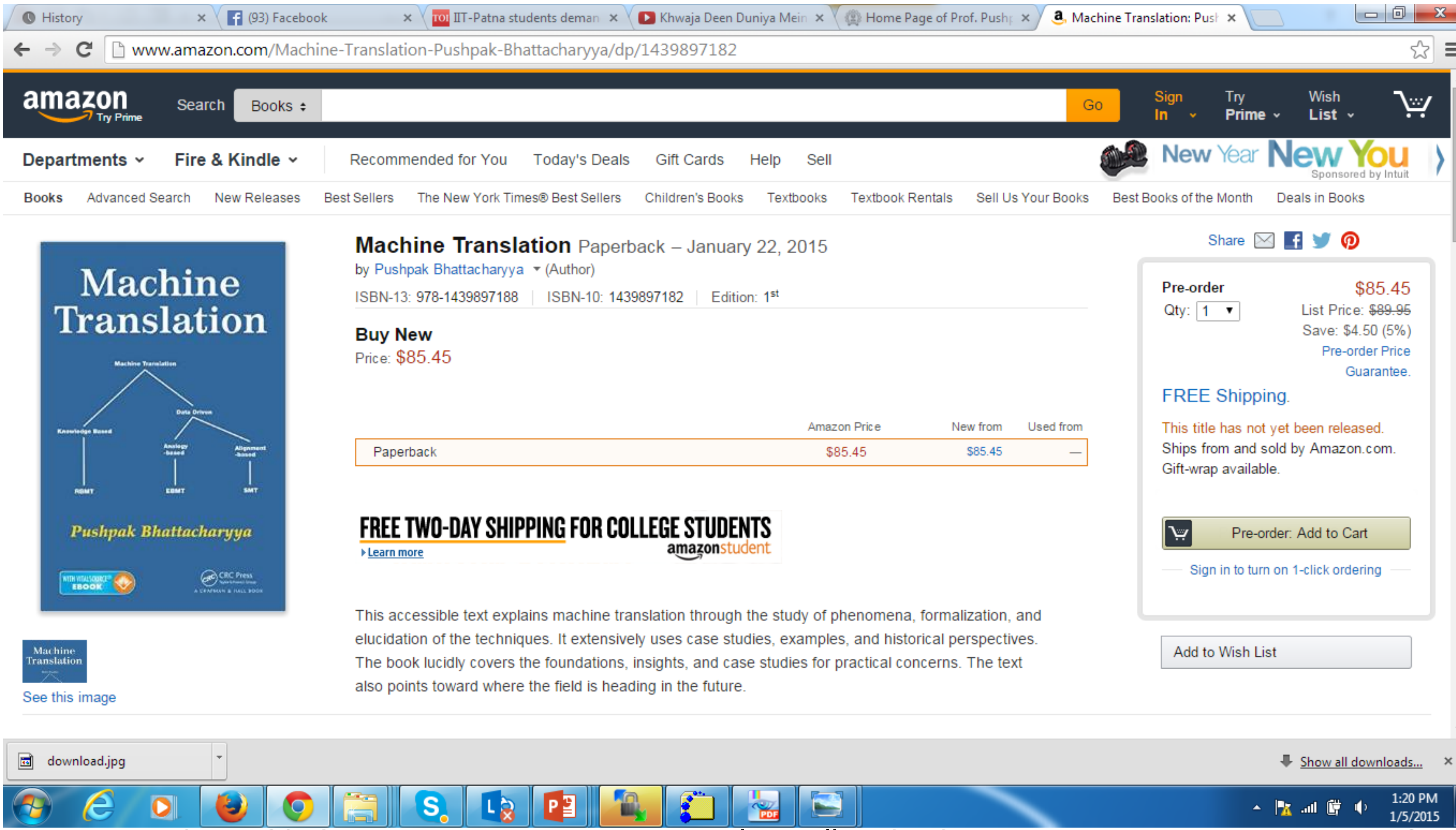
Indowordnet

Facts and figures- NLP@IITB

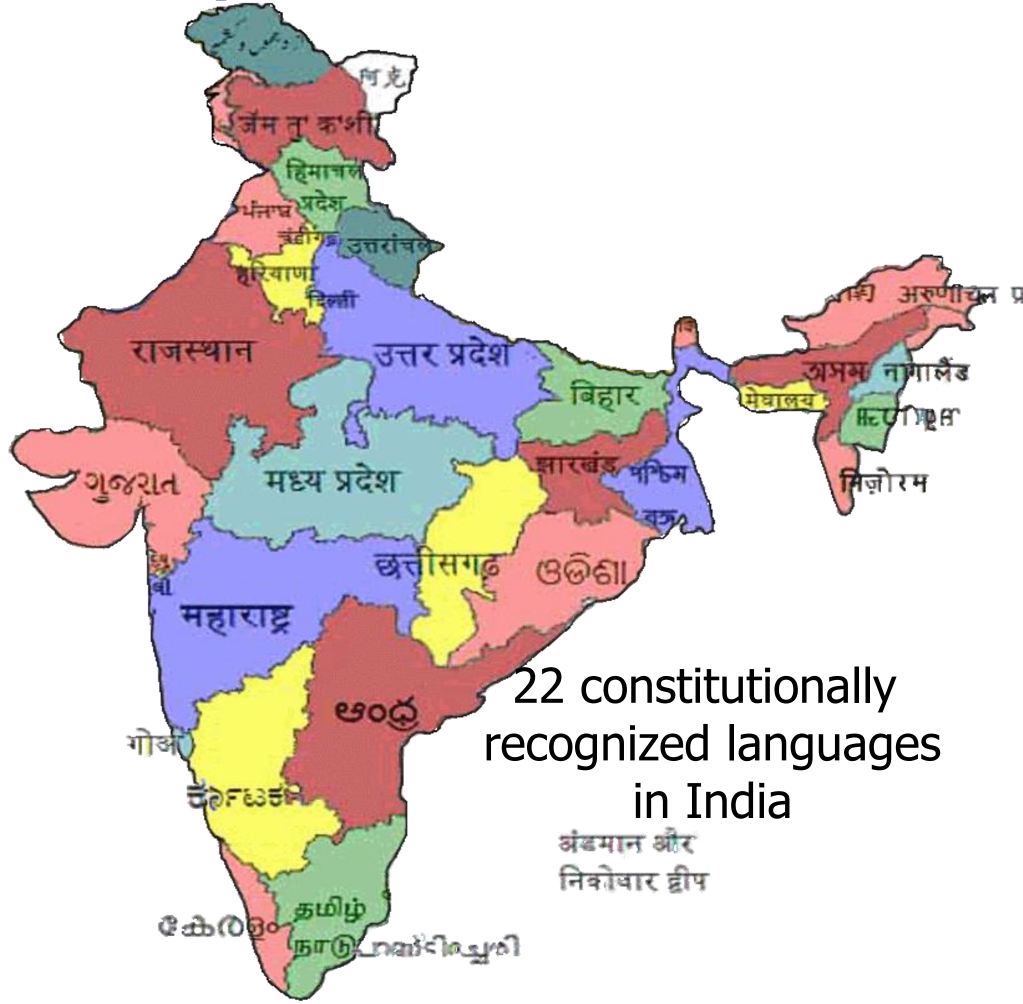
- 5 associated faculty: 3 CSE + 2 HSS
- PhD students: graduated-13; ongoing-17
- M.Tech students: graduated-115; ongoing-14
- Publications in highly visible fora: ACL, COLING, WWW, ECML, EMNLP
- Sponsorship: Ministry of IT, TCS, Yahoo, IBM, Microsoft, HP Labs, Xerox, AOL
- Technology developed used by major search engine companies of the world
- Major international conferences: COLING 2012, 8-16 DEC, 2012, at IIT Bombay- 700 participants

Book

http://www.amazon.com/Machine-Translation-Pushpak-Bhattacharyya/dp/1439897182



a multilingual world, A Multilingual country



22 constitutionally recognized languages in India

>1500 languages spoken in India

Pubs: http://ww.cse.iitb.ac.in/~pb

Resources and tools:

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

Thank You