CS626: Speech, NLP and the Web

RNN, Seq2seq, Data Driven Machine Translation (SMT and NMT)

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week of 9th November, 2020

Vauquois Triangle

Kinds of MT Systems

(point of entry from source to the target text)

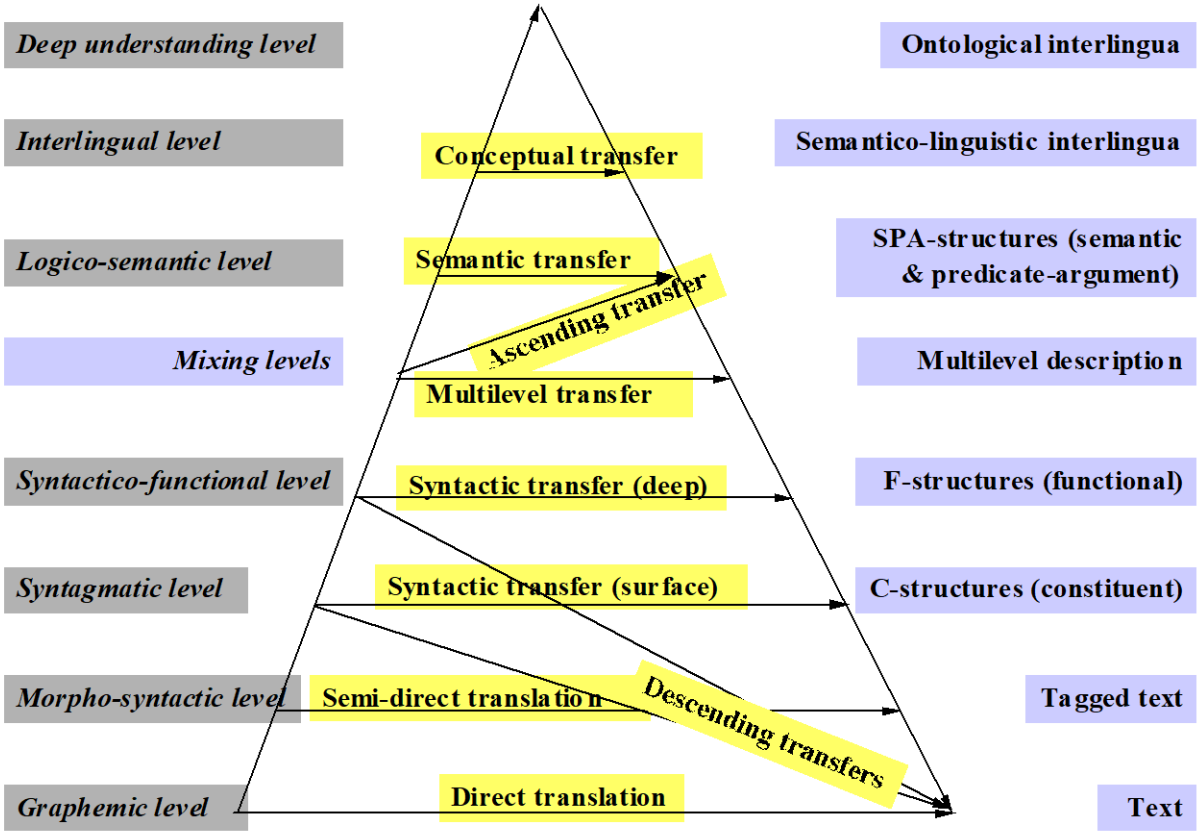
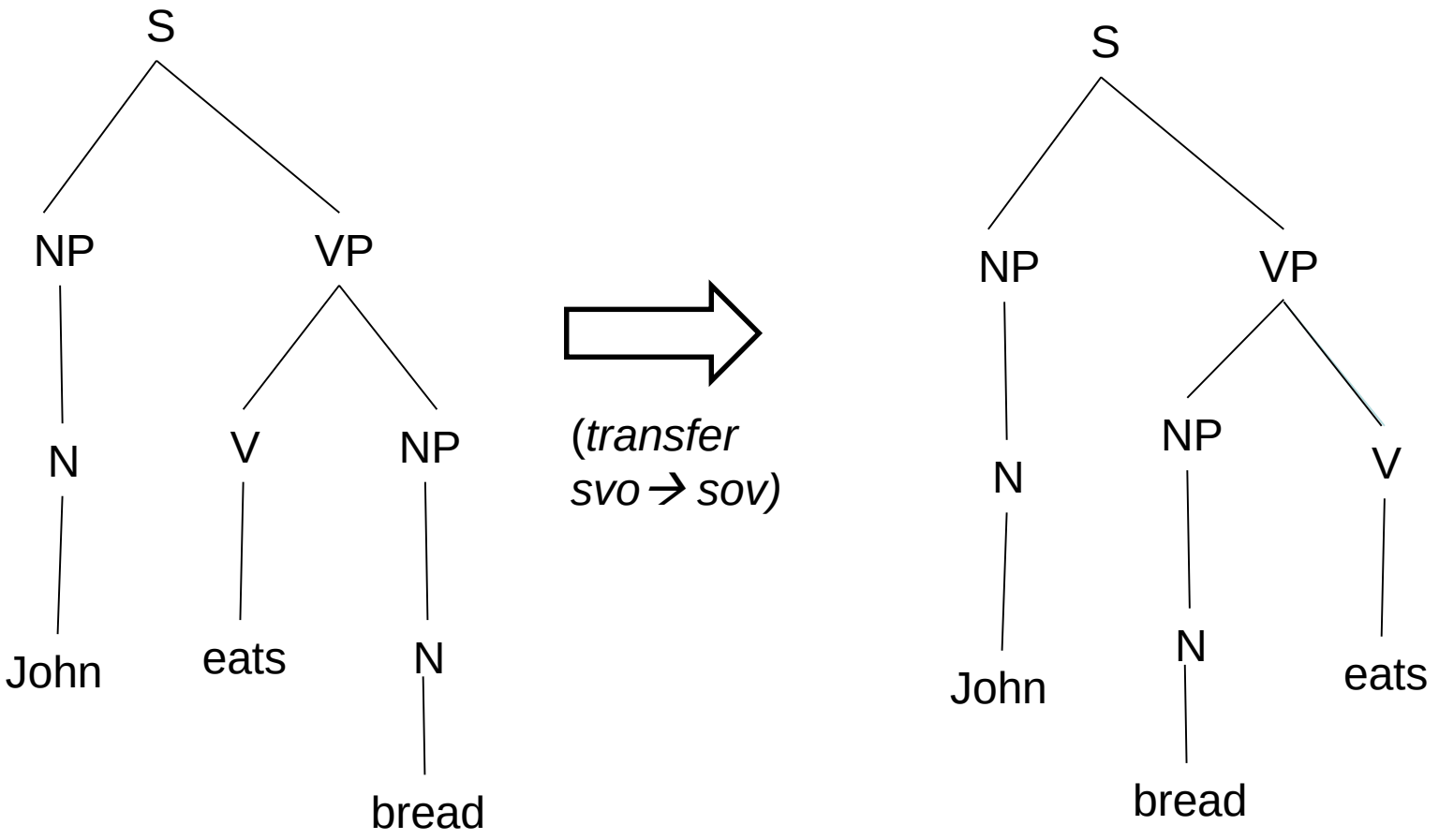


Illustration of transfer SVO→SOV



Fundamental processes in Machine Translation

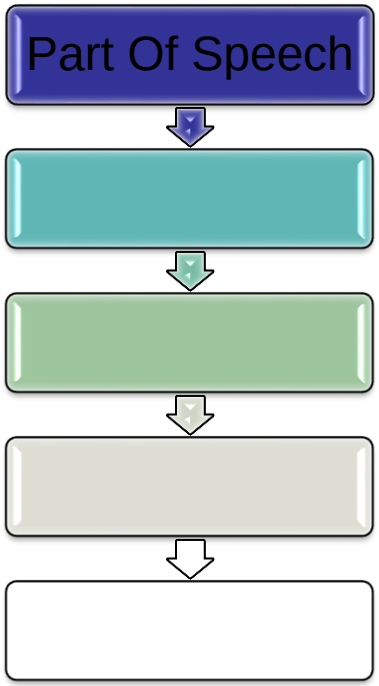
- **Analysis**
  - Analysis of the source language to represent the source language in more disambiguated form
    - Morphological segmentation, POS tagging, chunking, parsing, discourse resolution, pragmatics etc.
- **Transfer**
  - Knowledge transfer from one language to another
  - Example: SOV to SVO conversion
- **Generation**
  - Generate the final target sentence
  - Final output is text, intermediate representations can include F-structures, C-structures, tagged text etc.

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

Noun or Verb

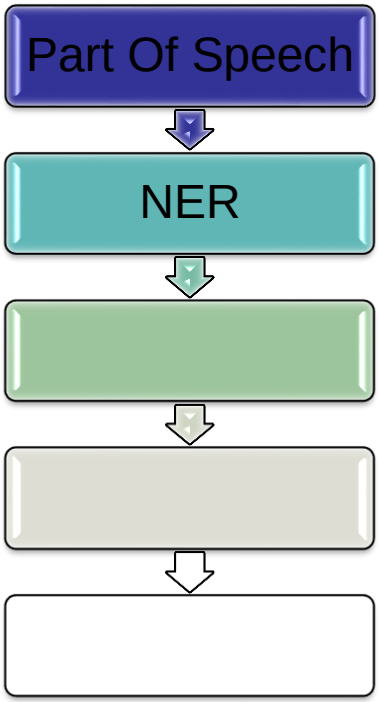
ISSUES



Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

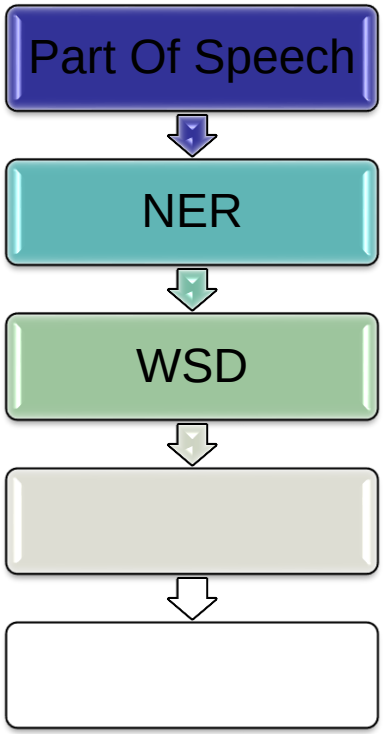


John is the name of a PERSON

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES



Financial bank or River bank

Issues to handle

Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

- Part Of Speech
- NER
- WSD
- Co-reference
-

"it" → "bank" .

Issues to handle

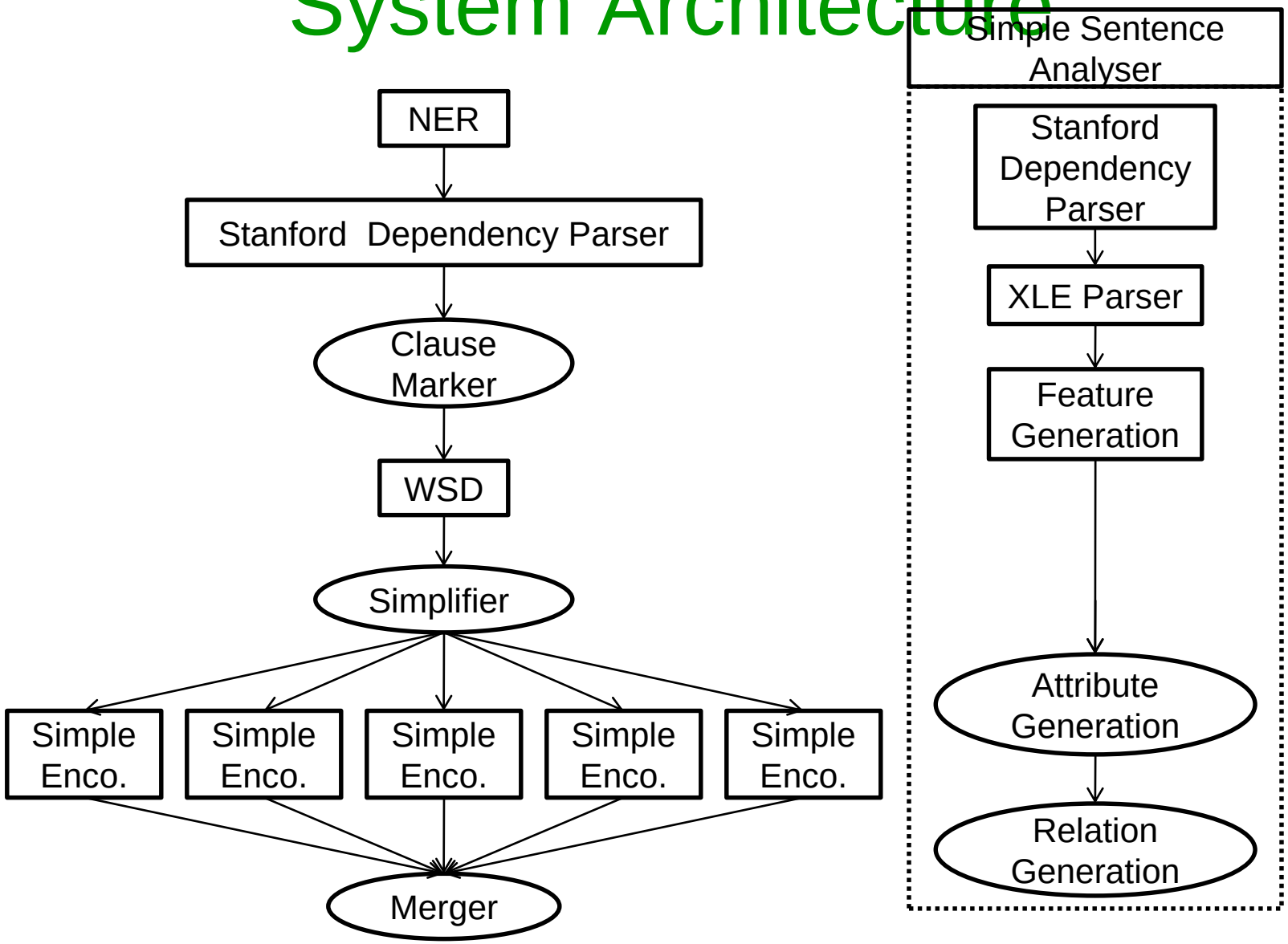
Sentence: *I went with my friend, John, to the bank to withdraw some money but was disappointed to find it closed.*

ISSUES

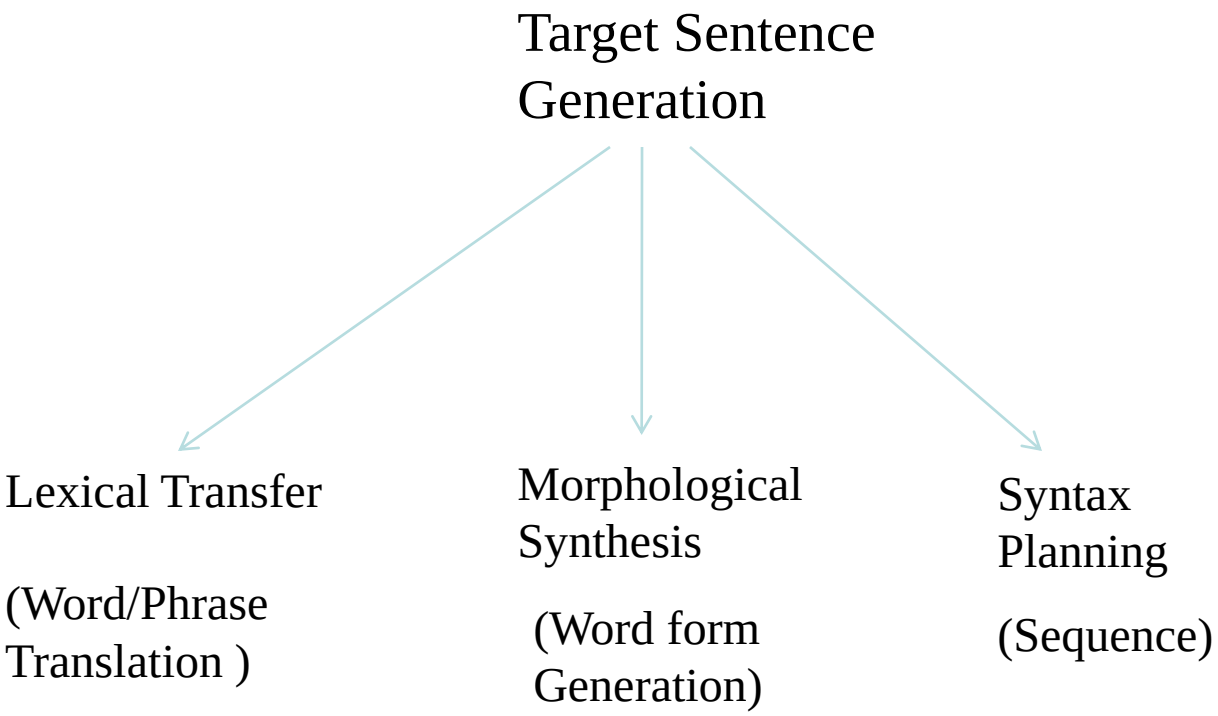
- Part Of Speech
- NER
- WSD
- Co-reference
- Subject Drop

Pro drop
(subject “I”)

System Architecture

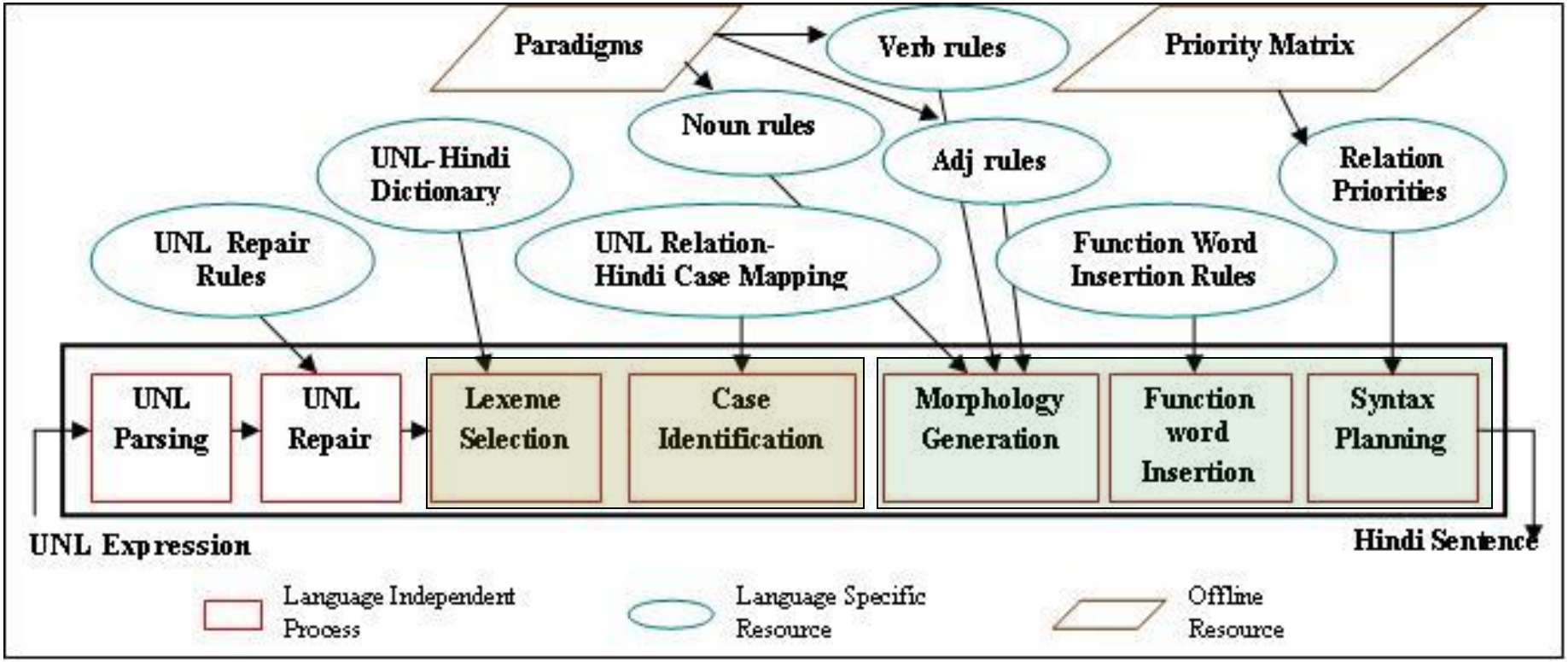


Target Sentence Generation from interlingua

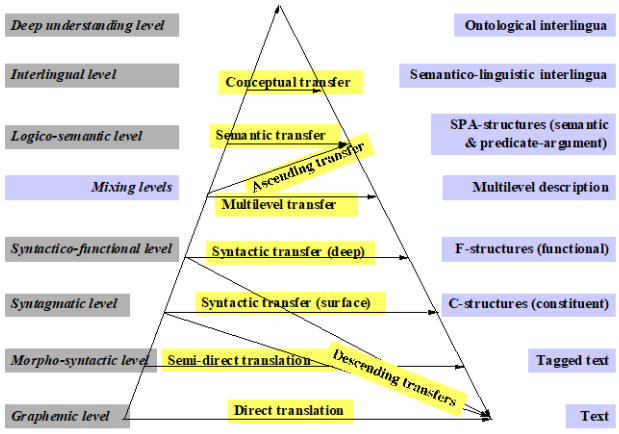


Generation Architecture

Deconversion = Transfer + Generation



Statistical Machine Translation



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.

Hindi-English data

- [DhotA huM] “I carry”
- [DhoegA] “He will carry”
- [DhotA hAi] “He carries”
- [Dhote hAi] “They carry”
- [chalAtA huM] “I drive”
- [tErte hEM] “They swim”

Bangla-English data

- [bai] “I carry”
- [baibe] “He will carry”
- [bay] “He carries”
- [bay] “They carry”
- [chAlAi] “I drive”
- [sAMtrAy] “They swim”

To translate … (repeated)

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

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

SMT: Language Model

- To detect *good* English sentences
- Probability of an English sentence $w_1 w_2 \ldots\ldots w_n$ can be written as

  $Pr(w_1 w_2 \ldots\ldots w_n) = Pr(w_1) * Pr(w_2/w_1) * \ldots * Pr(w_n/w_1\, w_2 \ldots w_{n-1})$
- Here $Pr(w_n/w_1\, w_2 \ldots w_{n-1})$ is the probability that word w_n follows word string $w_1\, w_2 \ldots w_{n-1}$.
  - N-gram model probability
- Trigram model probability calculation

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

SMT: Translation Model

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

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

– Sentences ⟵ Sentence level

to find pair(e,f) for all sentences

- Introduce a hidden variable $\boldsymbol{a}$, that represents alignments between the individual words in the sentence pair

$$\Pr(\boldsymbol{f}|\boldsymbol{e}) = \sum_{\boldsymbol{a}} \Pr(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e})$$

⟵ Word level

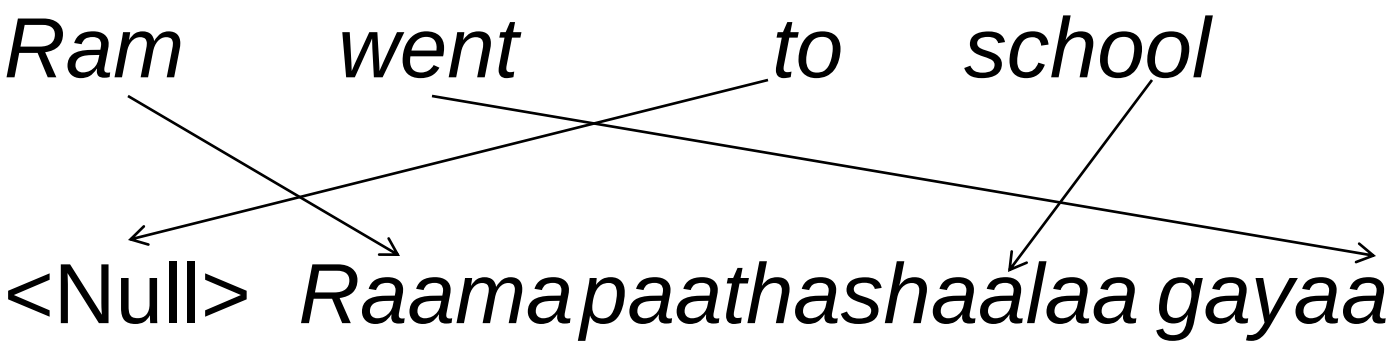
Alignment

- If the string, $e = e_1^l = e_1\ e_2\ \ldots e_l$, has l words, and the string, $f = f_1^m = f_1 f_2 \ldots f_m$, has m words,
- then the alignment, a, can be represented by a series, $\boldsymbol{a_1^m = a_1 a_2 \ldots a_m}$, of m values, each between 0 and l such that if the word in position j of the f-string is connected to the word in position i of the e-string, then
  - $\boldsymbol{a_j = i}$, and
  - if it is not connected to any English word, then $\boldsymbol{a_j = 0}$

Example of alignment

English: *Ram went to school*

Hindi: *Raama paathashaalaa gayaa*



Translation Model: Exact expression

$$\Pr(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = \Pr(m|\boldsymbol{e}) \prod_{j=1}^{m} \Pr\left(a_j \left| a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}\right.\right) \Pr\left(f_j \left| a_1^{j}, f_1^{j-1}, m, \boldsymbol{e}\right.\right)$$

Choose the length of foreign language string given *e*

Choose alignment given *e* and *m*

Choose the identity of foreign word given *e, m, a*

- Five models for estimating parameters in the expression [2]
- Model-1, Model-2, Model-3, Model-4, Model-5

Proof of Translation Model: Exact expression

$$\Pr(f \mid e) = \sum_{a} \Pr(f, a \mid e) \quad ; \text{ marginalization}$$

$$\Pr(f, a \mid e) = \sum_{m} \Pr(f, a, m \mid e) \quad ; \text{ marginalization}$$

$$\Pr(f, a, m \mid e) = \sum_{m} \Pr(m \mid e) \Pr(f, a \mid m, e)$$

$$= \sum_{m} \Pr(m \mid e) \Pr(f, a \mid m, e)$$

$$= \sum_{m} \Pr(m \mid e) \prod_{j=1}^{m} \Pr(f_j, a_j \mid a_1^{j-1}, f_1^{j-1}, m, e)$$

$$= \sum_{m} \Pr(m \mid e) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, e) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, e)$$

m is fixed for a particular *f*, hence

$$\Pr(f, a, m \mid e) = \Pr(m \mid e) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, e) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, e)$$

Alignment

Fundamental and ubiquitous

- Spell checking
- Translation
- Transliteration
- Speech to text
- Text to speeh

EM for word alignment from sentence alignment: example

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

The counts in IBM Model 1

Works by maximizing *P(f|e)* over the entire corpus

For IBM Model 1, we get the following relationship:

$$c(w^f \mid w^e; f, e) = \frac{t(w^f \mid w^e)}{t(w^f \mid w^{e_0}) + \square + t(w^f \mid w^{e_l})} \square \ldots \square$$

$c(w^f \mid w^e; f, e)$ is the fractional count of the alignment of w^f with w^e in f and e

$t(w^f \mid w^e)$ is the probability of w^f being the translation of w^e

$\square$ is the count of w^f in f

$\square$ is the count of w^e in e

Example of expected count

$C[a \leftrightarrow w;\ (a\ b) \leftrightarrow (w\ x)]$

$$= \frac{t(a \leftrightarrow w)}{t(a \leftrightarrow w)+t(a \leftrightarrow x)} \times \#(a \text{ in } \text{`a b'}) \times \#(w \text{ in } \text{`w x'})$$

$$= \frac{1/4}{1/4+1/4} \times 1 \times 1 = 1/2$$

"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 probability: example

$$t_{revised}(a \leftrightarrow w)$$

$$= \frac{1/2}{(1/2+1/2+0+0)_{(a\ b)\leftrightarrow(w\ x)} + (0+0+0+0)_{(b\ c\ d)\leftrightarrow(x\ y\ z)}}$$

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 of EM based Alignment Expressions

$$V_E = \text{vocalbulary of language } L_1 \text{ (Say English)}$$

$$V_F = \text{vocabulary of language } L_2 \text{ (Say Hindi)}$$

E^1 *what is in a name ?*
नाम में क्या है?
naam meM kya hai ?

F^1 *name in what is ?*
what is in a name ?

That which we call rose, by any other name will smell as sweet.

E^2 *जिसे हम गुलाब कहते हैं, और भी किसी नाम से उसकी कुशबू सामान मीठा होगी*

F^2 *Jise hum gulab kahte hai, aur bhi kisi naam se uski khushbu samaan mitha hogii*
That which we rose say , any other name by its smell as sweet
That which we call rose, by any other name will smell as sweet.

Vocabulary mapping

Vocabulary

V_E	V_F
what , is , in, a , name , that, which, we , call ,rose, by, any, other, will, smell, as, sweet	naam, meM, kya, hai, jise, hum, gulab, kahte, hai, aur, bhi, kisi, bhi, uski, khushbu, saman, mitha, hogii

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)

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

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

Combinatorial considerations

Example

E2.1: Peter went to school early

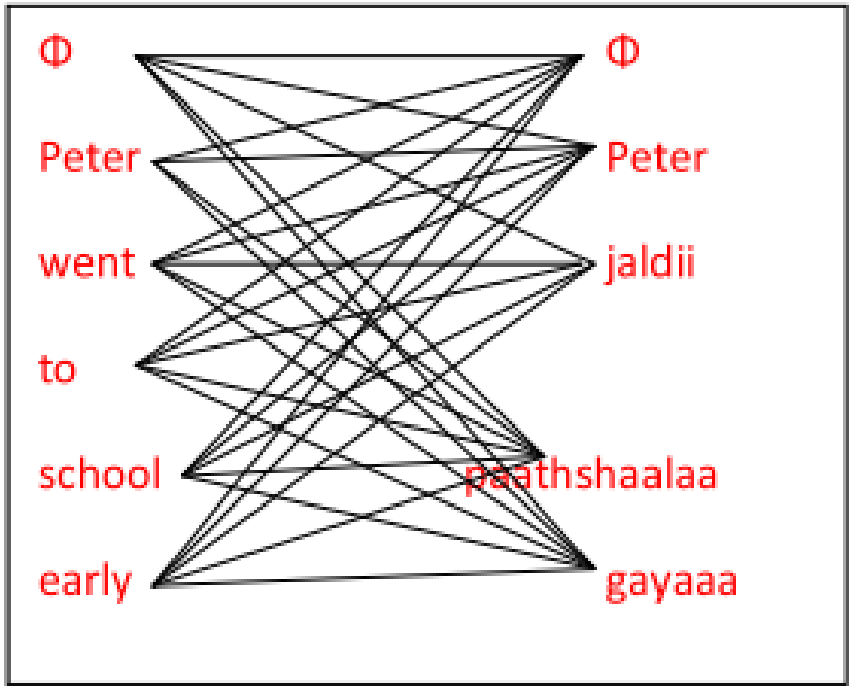
H2.1:पीटर जल्दी पाठशाला गया

T2.1: piitar jaldii paathshaalaa gayaa

G2.1: Peter early school went

Non English text

All possible alignments



First fundamental requirement of SMT

Alignment requires evidence of:

- firstly, a translation pair to introduce the POSSIBILITY of a mapping.
- then, another pair to establish with CERTAINTY the mapping

For the “certainty”

- We have a translation pair containing alignment candidates and **none** of the other words in the translation pair

OR

- We have a translation pair containing **all** words in the translation pair, except the alignment candidates

Therefore…

- *If M valid bilingual mappings exist in a translation pair then an additional M-1 pairs of translations will decide these mappings with certainty.*

Rough estimate of data requirement

- SMT system between two languages L_1 and L_2
- Assume no a-priori linguistic or world knowledge, *i.e.*, no meanings or grammatical properties of any words, phrases or sentences
- Each language has a vocabulary of 100,000 words
- can give rise to about 500,000 word forms, through various morphological processes, assuming, each word appearing in 5 different forms, on the average
  - For example, the word ‘go’ appearing in ‘go’, ‘going’, ‘went’ and ‘gone’.

Reasons for mapping to multiple words

- Synonymy on the target side (e.g., "*to go*" in English translating to "*jaanaa*", "*gaman karnaa*", "*chalnaa*" *etc.* in Hindi), a phenomenon called lexical choice or register
- polysemy on the source side (e.g., "*to go*" translating to "*ho jaanaa*" as in "*her face went red in anger*"→"*usakaa cheharaa gusse se laal ho gayaa*")
- syncretism ("went" translating to "*gayaa*", "*gayii*", or "*gaye*"). Masculine Gender, 1st or 3rd person, singular number, past tense, non-progressive aspect, declarative mood

Estimate of corpora requirement

- Assume that on an average a sentence is 10 words long.
- ➔ an additional 9 translation pairs for getting at one of the 5 mappings
- ➔ 10 sentences per mapping per word
- ➔ a first approximation puts the data requirement at 5 X 10 X 500000= 25 million parallel sentences
- Estimate is not wide off the mark
- Successful SMT systems like Google and Bing reportedly use 100s of millions of translation pairs.

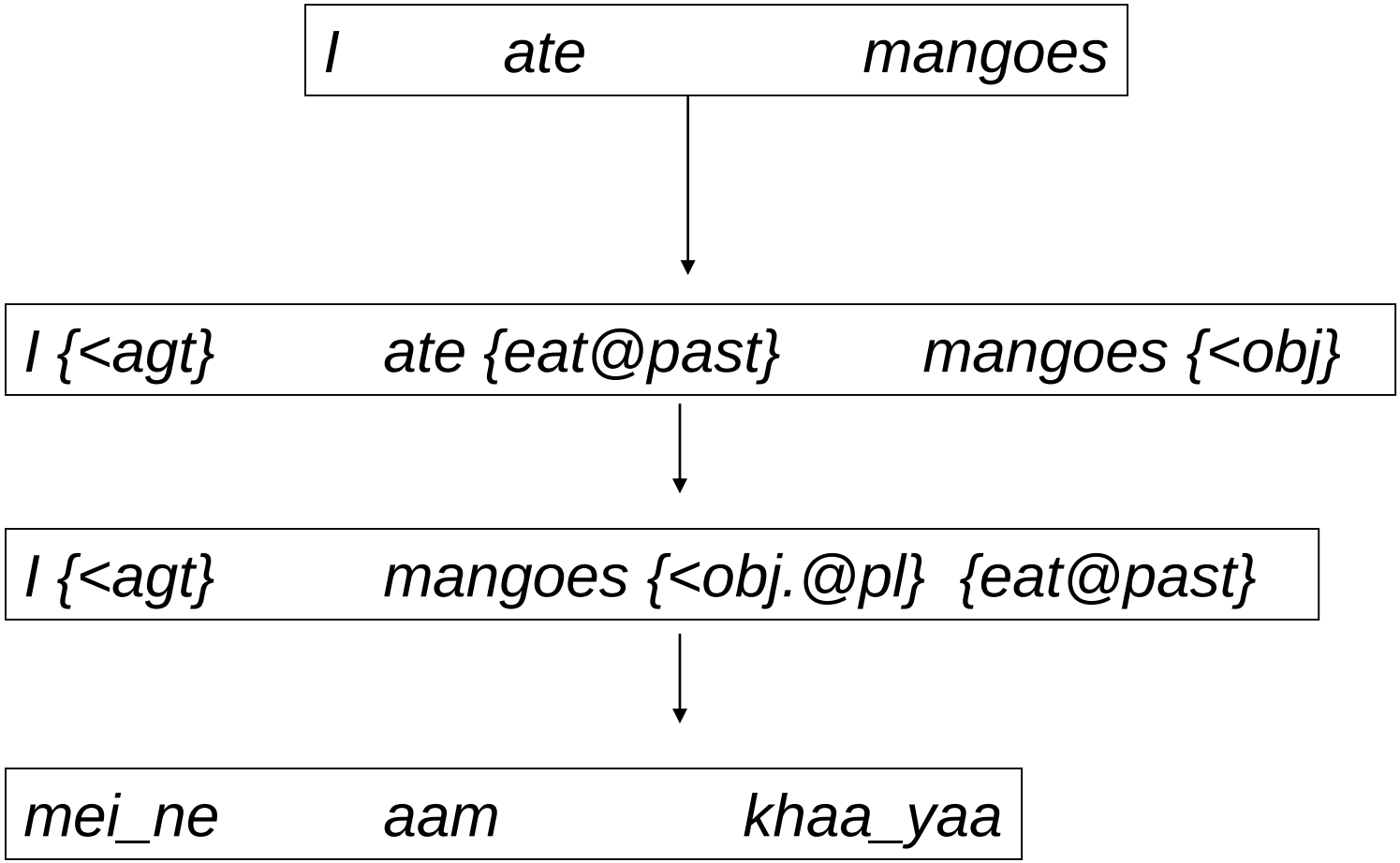
Our work on factor based SMT

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh and Pushpak Bhattacharyya, *Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, **ACL-IJCNLP 2009**, Singapore, August, 2009.*

Case Marker and Morphology crucial in E-H MT

- Order of magnitiude facelift in Fluency and fidelity
- Determined by the combination of suffixes and semantic relations on the English side
- Augment the aligned corpus of the two languages, with the correspondence of English suffixes and semantic relations with Hindi suffixes and case markers

Semantic relations+Suffixes→Case Markers+inflections

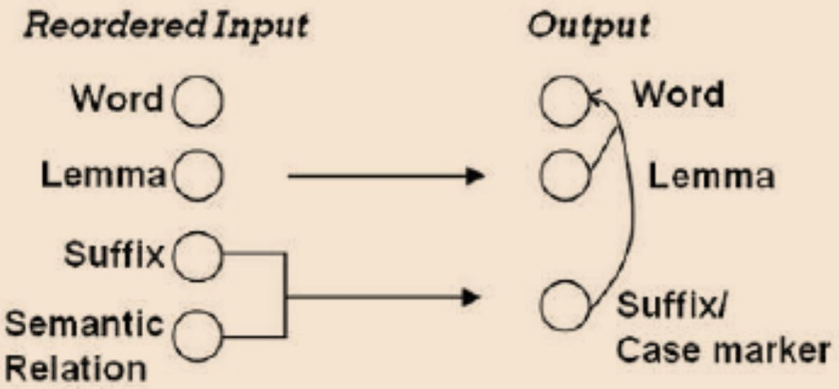


Our Approach

- Factored model (Koehn and Hoang, 2007) with the following translation factor:
  - **suffix + semantic relation → case marker/suffix**

- Experiments with the following relations:
  - Dependency relations from the stanford parser
  - Deeper semantic roles from Universal Networking Language (UNL)

Our Factorization



1. a lemma to lemma translation factor (boy → लडक् (*ladak*))
2. a suffix + semantic relation to suffix/case marker factor (-s + subj → ए (*e*))
3. a lemma + suffix to surface form generation factor (लडक् + ए (*ladak* + *e*) → लडके (*ladake*))

Experiments

Corpus Statistics

	#sentences	#words
Training	12868	316508
Tuning	600	15279
Test	400	8557

Results: The impact of suffix and semantic factors

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

Results: The impact of reordering and semantic relations

Model	Reordering	BLEU	NIST
surface	distortion	24.42	5.85
surface	lexicalized	28.75	6.19
surface	syntactic	31.57	6.40
lemma + suffix + stanford	syntactic	31.49	6.34

Subjective Evaluation: The impact of reordering and semantic relations

Model	Reordering	Fluency	Adequacy	#errors
surface	lexicalized	2.14	2.26	2.16
surface	syntactic	2.6	2.71	1.79
lemma + suffix + stanford	syntactic	2.88	2.82	1.44

Impact of sentence length (F: Fluency; A:Adequacy; E:# Errors)

	Baseline			Reorder			Stanford		
	F	A	E	F	A	E	F	A	E
Small (<19 words)	2.63	2.84	1.30	3.30	3.52	0.74	3.66	3.75	0.62
Medium (20-34 words)	1.92	2.00	2.23	2.32	2.43	2.05	2.62	2.46	1.74
Large (>34 words)	1.62	1.69	4.00	1.86	1.73	3.36	1.86	1.86	2.82

A feel for the improvement-baseline

Input: Inland waterway is one of the most popular picnic spots in Alappuzha.

Baseline: में एक अन्त:स्थलीय जलमार्ग के सबसे प्रसिद्ध पिकनिक स्थल में जलों में दौडती है

men eka antahsthaliiya jalamaarga ke sabase prasiddha pikanika sthala men jalon men daudatii hai

gloss: in a waterway of most popular picnic spot in waters runs.

A feel for the improvement-reorder

Reorder: अन्त:स्थलीय जलमार्ग आलपुया के सबसे प्रसिद्ध पिकनिक स्थल में से एक है

antahsthaliiya jalamaarga aalapuzaa ke sabase prasiddha pikanika sthala men se eka hai

A feel for the improvement-Semantic relation

Semantic: अन्त:स्थलीय जलमार्ग आलपुया के सबसे प्रसिद्ध पिकनिक स्थलों में से एक है

antahsthaliiya jalamaarga aalapuzaa ke sabase prasiddha pikanika sthalon men se eka hai

gloss: waterway Alappuzha of most popular picnic spots of one is

A recent study

PAN Indian SMT

Pan-Indian Language SMT

http://www.cfilt.iitb.ac.in/indic-translator

- SMT systems between 11 languages
  - 7 Indo-Aryan: Hindi, Gujarati, Bengali, Oriya, Punjabi, Marathi, Konkani
  - 3 Dravidian languages: Malayalam, Tamil, Telugu
  - English
- Corpus
  - Indian Language Corpora Initiative (ILCI) Corpus
  - Tourism and Health Domains
  - 50,000 parallel sentences
- Evaluation with BLEU
  - METEOR scores also show high correlation with BLEU

SMT Systems Trained

- **Phrase-based** (PBSMT) baseline system (S1)
- E-IL PBSMT with **Source side reordering rules** *(Ramanathan et al., 2008)* (S2)
- E-IL PBSMT with **Source side reordering rules** *(Patel et al., 2013)* (S3)
- IL-IL PBSMT with **transliteration post-editing** (S4)

Natural Partitioning of SMT systems

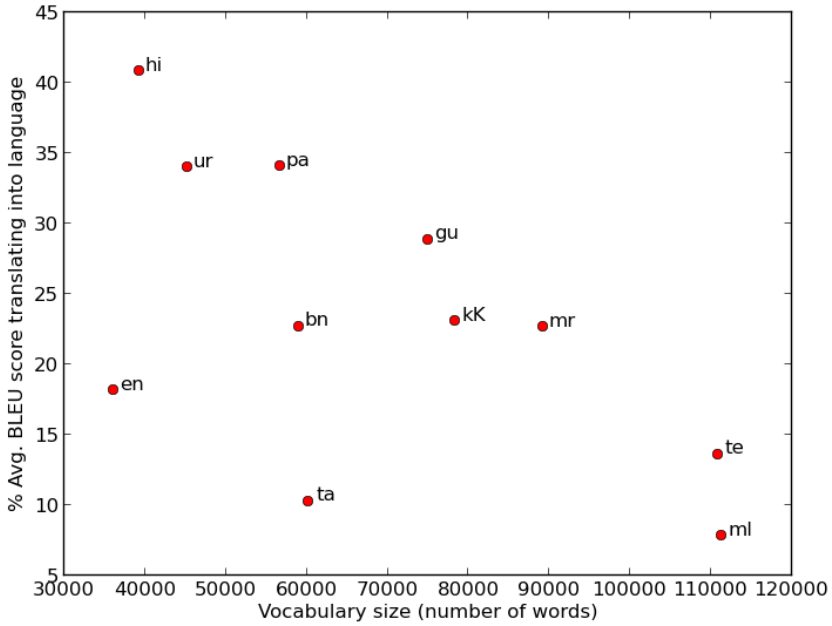
	hi	ur	pa	bn	gu	mr	kK	ta	te	ml	en
hi		61.28	68.21	34.96	51.31	39.12	37.81	14.43	21.38	10.98	29.23
ur	61.42		52.02	29.59	39.00	27.57	28.29	11.95	16.61	8.65	22.46
pa	73.31	56.00		29.89	43.85	30.87	30.72	10.75	18.81	9.11	23.83
bn	37.69	32.08	31.38		28.14	22.09	23.47	10.94	13.40	8.10	18.76
gu	55.66	44.12	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)

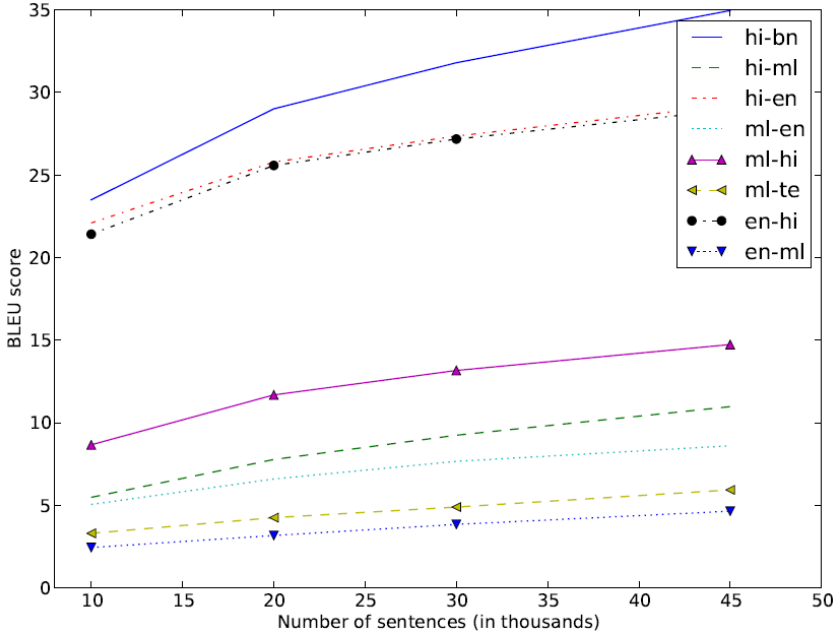
- **Clear partitioning of translation pairs by language family pairs**, based on translation accuracy.
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult

The Challenge of Morphology

Morphological complexity vs BLEU



Training Corpus size vs BLEU



Vocabulary size is a proxy for morphological complexity
*Note: For Tamil, a smaller corpus was used for computing vocab size

- Translation accuracy decreases with increasing morphology
- Even if training corpus is increased, commensurate improvement in translation accuracy is not seen for morphologically rich languages
- ***Handling morphology in SMT is critical***

Common Divergences, Shared Solutions

System	hi	ur	pa	bn	gu	mr	kK	ta	te	ml
Baseline PBSMT	28.94	22.96	22.33	15.33	15.44	12.11	13.66	6.43	6.55	4.65
Source Reordering (Generic)	31.41	24.85	24.56	15.89	17.38	13.42	14.55	7.84	8.23	4.95
Source Reordering (Hindi-adapted)	33.54	26.67	26.23	17.86	19.06	14.15	15.56	7.96	8.37	5.30

Comparison of source reordering methods for E-IL SMT - % BLEU scores (S1,S2,S3)

- All Indian languages have similar word order
- The same structural divergence between English and Indian languages *SOV<->SVO, etc.*
- **Common source side reordering rules** improve E-IL translation by 11.4% (generic) and 18.6% (Hindi-adapted)
- ***Common divergences can be handled in a common framework in SMT systems*** (This idea has been used for knowledge based MT systems e.g. *Anglabharati*)

Harnessing Shared Characteristics

	hi	ur	pa	bn	gu	mr	kK	ta	te	ml
hi		61.28	64.85	35.49	52.98	39.12	37.81	14.52	21.68	11.07
ur	61.42		52.02	29.59	39.00	27.57	28.29	11.95	16.61	8.65
pa	74.14	56.00		30.05	44.39	31.46	30.99	10.77	18.96	9.12
bn	38.17	32.08	31.54		28.73	22.60	23.79	10.97	13.52	8.17
gu	57.22	44.12	45.55	28.90		33.22	31.55	12.64	17.46	8.05
mr	45.11	32.60	30.97	24.09	33.48		27.81	10.80	13.12	7.68
kK	41.92	34.00	32.04	24.91	32.05	27.52		10.40	14.92	7.96
ta	20.54	18.12	15.57	13.25	16.57	11.64	11.94		8.57	6.40
te	29.23	25.07	25.67	16.68	21.20	15.19	17.43	8.71		6.77
ml	14.81	13.39	12.98	10.73	9.84	8.42	9.25	5.99	6.02	

PBSMT+ transliteration post-editing for E-IL SMT - % BLEU scores (S4)

- Out of Vocabulary words are transliterated in a post-editing step
- Done using a simple transliteration scheme which harnesses the common phonetic organization of Indic scripts
- Accuracy Improvements of 0.5 BLEU points with this simple approach
- ***Harnessing common characteristics can improve SMT output***

Cognition and Translation: Measuring Translation Difficulty

Abhijit Mishra and Pushpak Bhattacharyya, *Automatically Predicting Sentence Translation Difficulty*, **ACL 2013**, Sofia, Bulgaria, 4-9 August, 2013

Scenario

Sentences

- *John ate jam*
- *John ate jam made from apples*
- *John is in a jam*

Subjective notion of difficulty

- *Easy*
- *Moderate*
- *Difficult?*

Use behavioural data

- Use behavioural data to decipher strong AI algorithms
- Specifically,
  - For WSD by humans, see where the eye rests for clues
  - For the innate translation difficulty of sentences, see how the eye moves back and forth over the sentences

Eye-tracking

Image Courtesy: http://www.smashingmagazine.com/2007/10/09/30-usability-issues-to-be-aware-of/

Eye Tracking data

- **Gaze points** : Position of eye-gaze on the screen
- **Fixations** : A long stay of the gaze on a particular object on the screen.
  - Fixations have both Spatial (coordinates) and Temporal (duration) properties.
- **Saccade** : A very rapid movement of eye between the positions of rest.
- **Scanpath**: A path connecting a series of fixations.
- **Regression**: Revisiting a previously read segment

Controlling the experimental setup for eye-tracking

- Eye movement patterns influenced by factors like age, working proficiency, environmental distractions etc.
- Guidelines for eye tracking
  - Participants metadata (age, expertise, occupation) etc.
  - Performing a fresh calibration before each new experiment
  - Minimizing the head movement
  - Introduce adequate line spacing in the text and avoid scrolling
  - Carrying out the experiments in a relatively low light environment

Use of eye tracking

- Used extensively in Psychology
  - Mainly to study reading processes
  - Seminal work: Just, M.A. and Carpenter, P.A. (1980). A theory of reading: from eye fixations to comprehension. Psychological Review 87(4):329–354

- Used in flight simulators for pilot training

NLP and Eye Tracking research

- Kliegl (2011)- Predict word frequency and pattern from eye movements
- Doherty et. al (2010)- Eye-tracking as an automatic Machine Translation Evaluation Technique
- Stymne et al. (2012)- Eye-tracking as a tool for Machine Translation (MT) error analysis
- Dragsted (2010)- Co-ordination of reading and writing process during translation.

Relatively new and open research direction

Translation Difficulty Index (TDI)

- Motivation: route sentences to translators with right competence, as per difficulty of translating
  - On a crowdsourcing platform, e.g.
- TDI is a function of
  - *sentence length (l)*,
  - *degree of polysemy of constituent words (p)* and
  - *structural complexity (s)*

Contributor to TDI: length

- What is more difficult to translate?
  - *John eats jam*
    - vs.
  - *John eats jam made from apples*
    - *vs.*
  - *John eats jam made from apples grown in orchards*
    - vs.
  - *John eats bread made from apples grown in orchards on black soil*

Contributor to TDI: polysemy

- What is more difficult to translate?
  - *John is in a jam*
    - vs.
  - *John is in difficulty*

- *Jam* has 4 diverse senses, *difficulty* has 4 related senses

Contributor to TDI: structural complexity

- What is more difficult to translate?
  - *John is in a jam. His debt is huge. The lenders cause him to shy from them, every moment he sees them.*
    - vs.
  - *John is in a jam, caused by his huge debt, which forces him to shy from his lenders every moment he sees them.*

Measuring translation through Gaze data

- Translation difficulty indicated by
  - staying of eye on segments
  - Jumping back and forth between segments

Example:

- *The horse raced past the garden fell*

Measuring translation difficulty through Gaze data

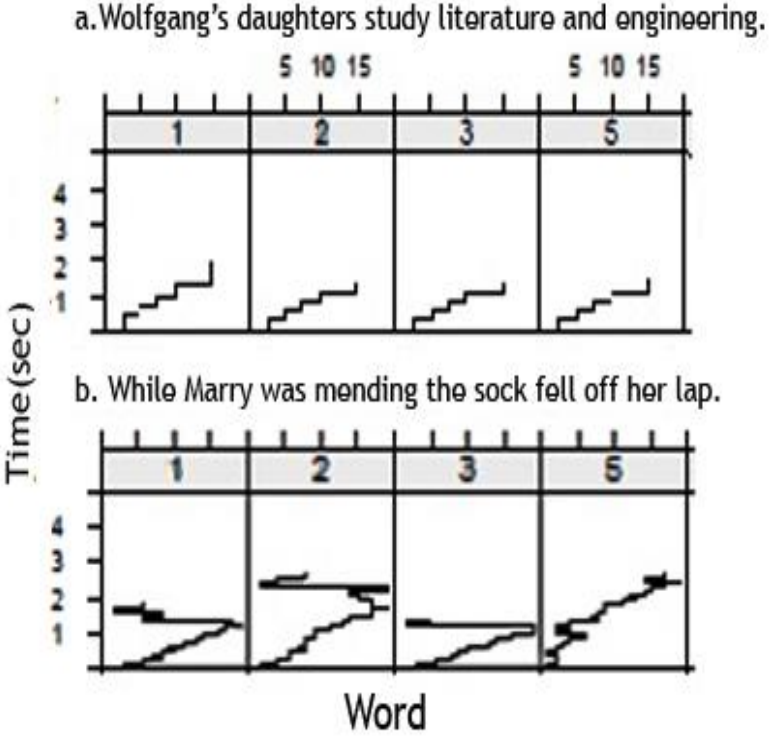
- Translation difficulty indicated by
  - staying of eye on segments
  - Jumping back and forth between segments

  Example:
    - *The horse raced past the garden fell*
    - *बगीचा के पास से दौडाया गया घोड़ा गिर गया*
    - *bagiichaa ke pas se doudaayaa gayaa ghodaa gir gayaa*

The translation process will complete the task till *garden*, and then backtrack, revise, restart and translate in a different way

Scanpaths: indicator of translation difficulty

- (Malsburg et. al. 2007)



- Sentence 2 is a clear case of "Garden pathing" which imposes cognitive load on participants and the prefer syntactic re-analysis.

Translog : A tool for recording Translation Process Data

- Translog (Carl, 2012) : A Windows based program
- Built with a purpose of recording gaze and key-stroke data during translation
- Can be used for other reading and writing related studies
- Using Translog, one can:
  - Create and Customize translation/reading and writing experiments involving eye-tracking and keystroke logging
  - Calibrate the eye-tracker
  - Replay and analyze the recorded log files
  - Manually correct errors in gaze recording

TPR Database

- The Translation Process Research (TPR) database (Carl, 2012) is a database containing behavioral data for translation activities
- Contains Gaze and Keystroke information for more than 450 experiments
- 40 different paragraphs are translated into 7 different languages from English by multiple translators
- At least 5 translators per language
- Source and target paragraphs are annotated with POS tags, lemmas, dependency relations etc
- Easy to use XML data format

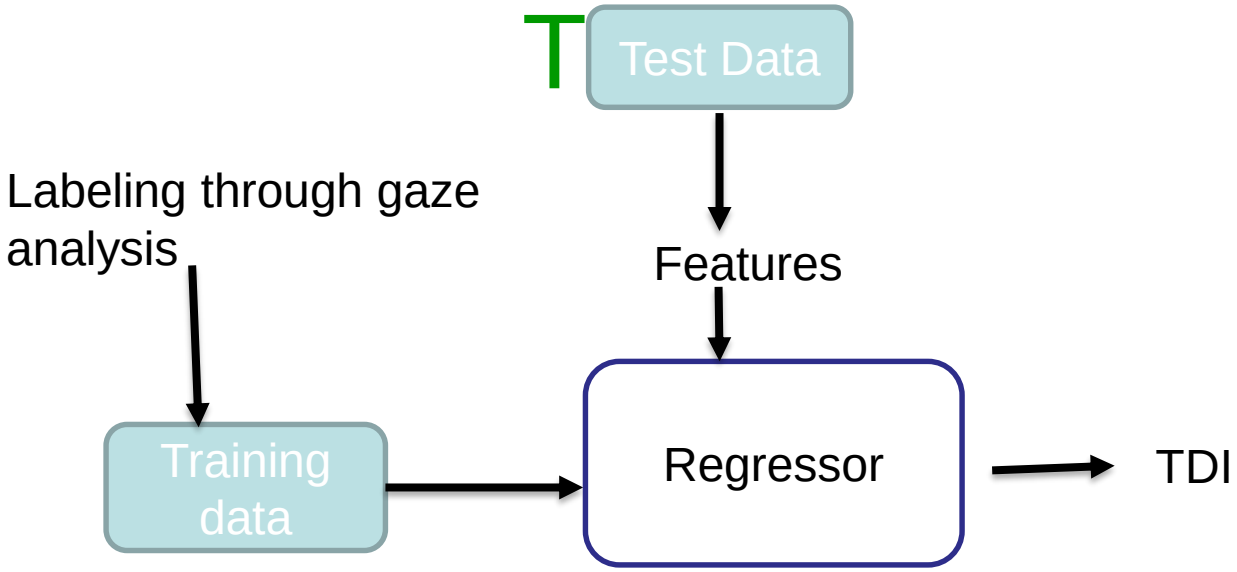
Experimental setup (1/2)

- Translators translate sentence by sentence typing to a text box
- The display screen is attached with a remote eye-tracker which
- constantly records the eye movement of the translator

Experimental setup (2/2)

- Extracted 20 different text categories from the data
- Each piece of text contains 5-10 sentences
- For each category we had at least 10 participants who translated the text into different target languages .

A predictive framework for T



- Direct annotation of TDI is fraught with subjectivity and ad-hocism.
- We use translator's gaze data as annotation to prepare training data.

Annotation of TDI (1/4)

- First approximation -> TDI equivalent to “time taken to translate”.
- However, time taken to translate may not be strongly related to translation difficulty.
  - It is difficult to know what fraction of the total time is spent on translation related thinking.
  - Sensitive to distractions from the environment.

Annotation of TDI (2/4)

- Instead of the "time taken to translate", consider "time for which translation related processing is carried out by the brain"
- This is called *Translation Processing Time*, given by:

$$T_p = T_{comp} + T_{gen}$$

- T_{comp} and T_{gen} are the comprehension of source text comprehension and target text generation respectively.

Annotation of TDI (3/4)

Humans spend time on what they see, and this "time" is correlated with the complexity of the information being processed

f- fixation, s- saccade, F_s- source, F_t- target

$$T_p = \sum_{f \in F_s} dur(f) + \sum_{s \in S_s} dur(s) +$$

Annotation of TDI (4/4)

- The measured TDI score is the T_p normalized over sentence length

$$TDI_{measured} = \frac{T_p}{sentence\_length}$$

Features

- **Length**: Word count of the sentences
- **Degree of Polysemy**: Sum of number of senses of each word in the WordNet normalized by length
- **Structural Complexity**: If the attachment units lie far from each other, the sentence has higher structural complexity. Lin (1996) defines it as the total length of dependency links in the dependency structure of the sentence.

Measured TDI for TPR database for 80 sentences.

Experiment and results

- Training data of 80 examples; 10-fold cross validation
- Features computed using Princeton WordNet and Stanford Dependency Parser
- Support Vector Regression technique (Joachims et al., 1999) along with different kernels
- Error analysis was done by Mean Squared Error estimate
- We also computed the correlation of the predicted TDI with the measured TDI.

Kernel(C=3.0)	MSE (%)	Correlation
Linear	20.64	0.69
Poly (Deg 2)	**12.88**	**0.81**
Poly (Deg 3)	13.35	0.78
Rbf (default)	13.32	0.73

Examples from the dataset

Example	L	DP	SC	TDI_O	TDI_P	Error
1. American Express recently announced a second round of job cuts.	10	10	1.8	0.24	0.23	4%
2. Sociology is a relatively new academic discipline.	7	6	3.7	0.49	0.53	8%

Summary

- Covered Interlingual based MT: the oldest approach to MT
- Covered SMT: the newest approach to MT
- Presented some recent study in the context of Indian Languages.

Summary

- SMT is the ruling paradigm
- But linguistic features can enhance performance, especially the factored based SMT with factors coming from interlingua
- Large scale effort sponsored by ministry of IT, TDIL program to create MT systems
- Parallel corpora creation is also going on in a consortium mode

Conclusions

- NLP has assumed great importance because of large amount of text in e-form
- Machine learning techniques are increasingly applied
- Highly relevant for India where multilinguality is way of life
- Machine Translation is more fundamental and ubiquitous than just mapping between two languages
- Utterance←→thought
- Speech to speech online translation

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

Resources and tools:

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