Day 6

Speaker: Pushpak Bhattacharyya, IIT Bombay

Topic: Machine Translation

Morphenes → The smallest part in a word having a meaning

Eg Inconceivable

In | conceive | able → Morphenes

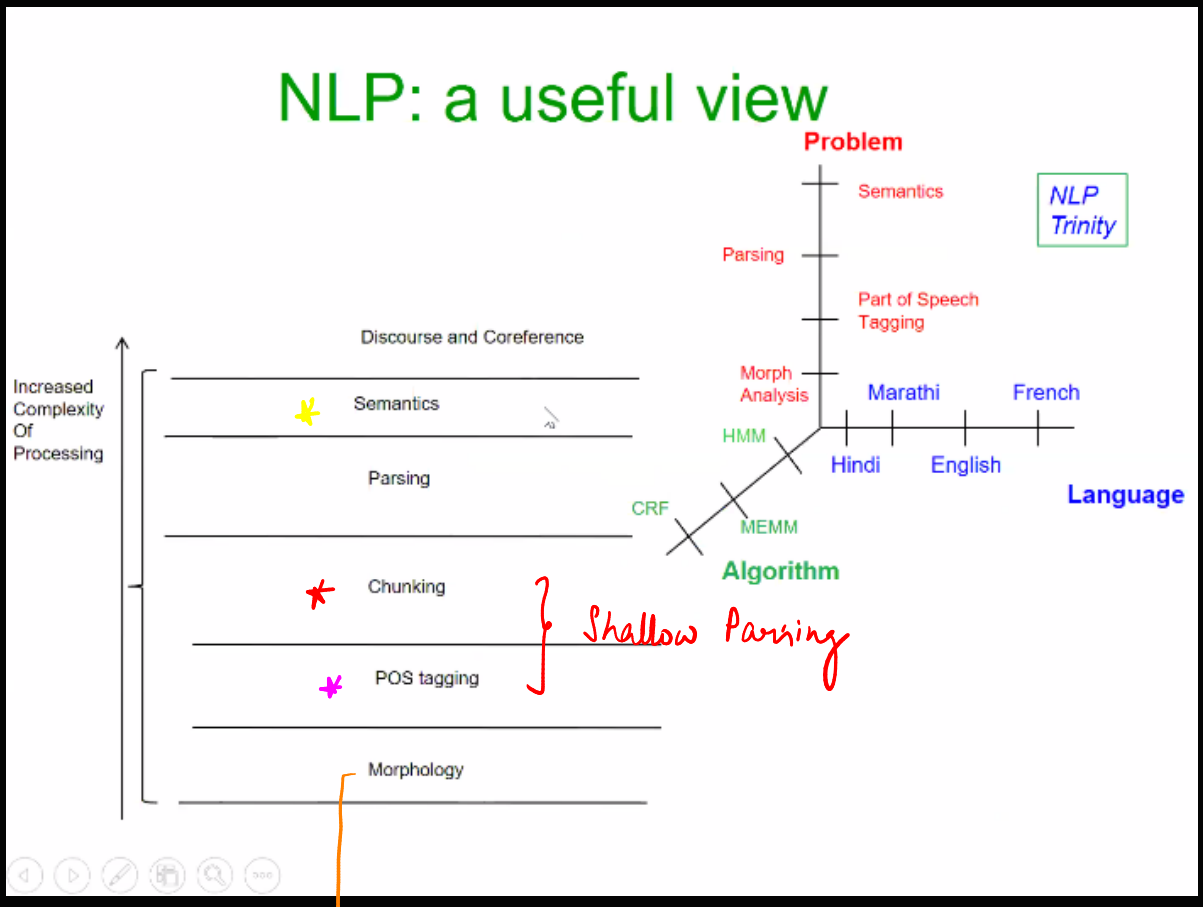
Languages

(1 word → 1 meaning)

Isolating ← Analytic

Synthetic → Agglutinative (glue many words together)

(1 word many meaning)



Breaking a word into its parts. → One category

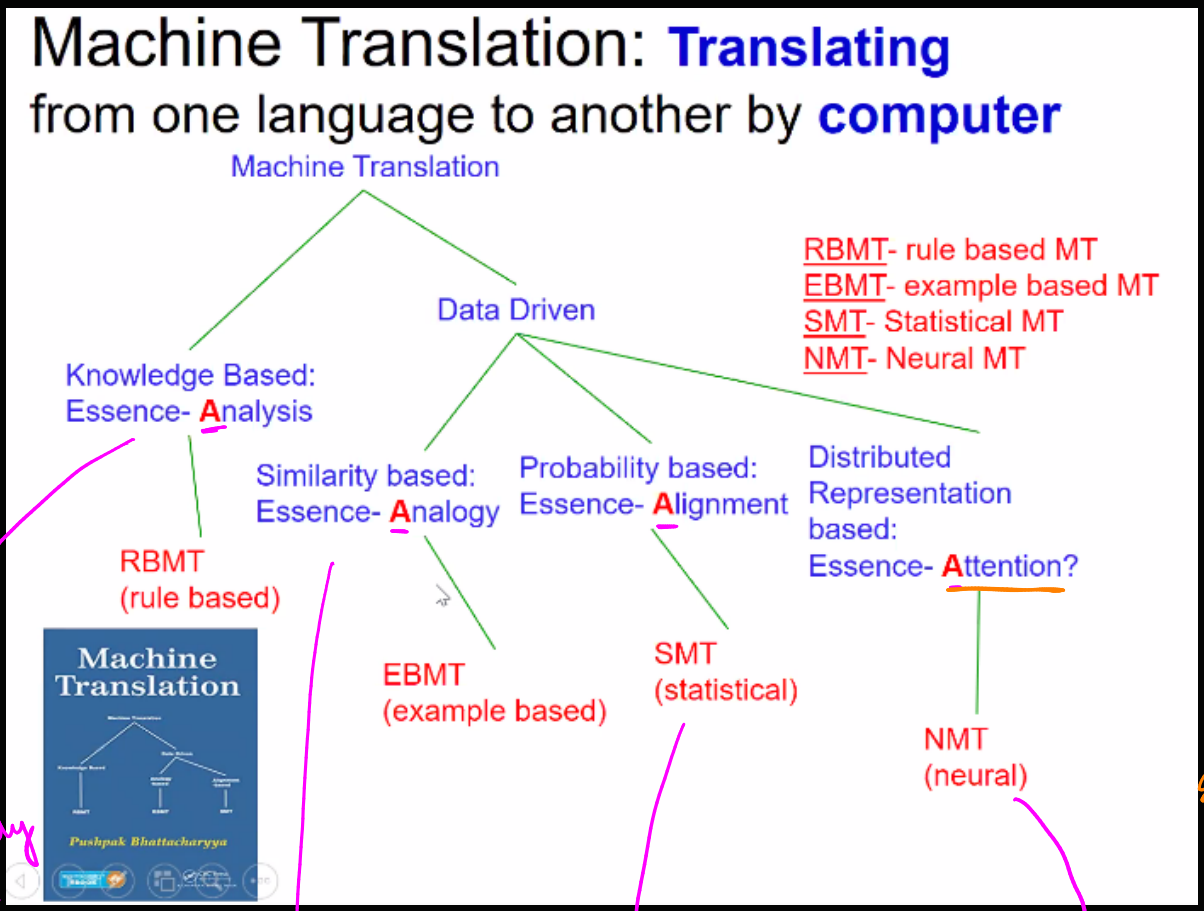
↓

Agglutination

(Stitching many words together)

Extreme Agglutination → Manipuri etc

* Part of Speech (POS Tagging) → Classification Problem
  For each part in a word or sentence → category is assigned
* Chunking (in case of small phrases)
* Semantics → Semantic Role Identification (Place, Time, etc)
  Sense of word
* Rule Based MT ~ Still Important.



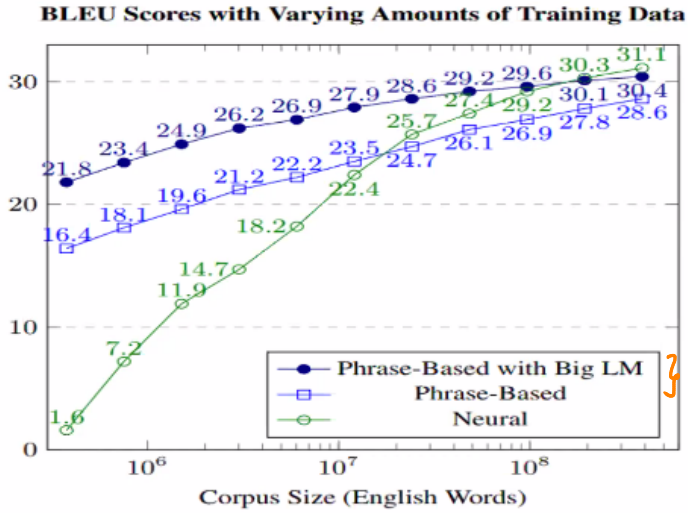
Source Sentence is analysed & converted to a form to directly generate target sentence
≃ 1960's & 1970's NLP stack is born
[Analysis needs to be good]

Can reuse (translation memory) Analogy

IBM → 1990's
Expectation Maximization
{Makes alignment betn words in a pair of sentences}
Source & Target Sentences

Encoder/Decoder & Transformers

Today's Ruling Paradigm: NMT which is data intensive



→ Statistical (SMT)

Philipp Koehn and Rebecca Knowles. 2017. *Six Challenges for Neural Machine Translation.* In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

As data ↑, Neural N/W method (Transformers) beats all other method

SMT etc.

But!!

* Neural → Lack of interpretability/explainability

Challenge of MT: Language Divergence

- 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)

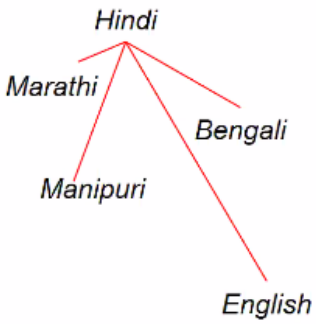
Understanding Divergences:

Different ways of expressing meaning

Indo-European

Tibeto-Burman

English:
This blanket is very soft
Hindi:
yaha kambal bahut naram hai
Bangla:
ei kambal ti khub naram <null>
Marathi:
haa kambal khup naram aahe
Manipuri:
kampor asi mon mon laui
blanket this soft soft is



Definitive Pronoun (DP) — This blanket is very soft — intensifier relationship } Meaning Graph.
Definitive relation — Stage Relationship

English → DP Subject Verb Object / Hindi → DP Subject Object Verb Copular

Bangla → DP Subject Object (→ ti?) Verb Copula?? / Marathi = " (same) /

Manipuri → Object (This) (DP) Object² Verb Copular (No intensifier (very))

* Hindi, Bangla & Marathi → almost 1 to 1 map
* Bangla clanifier → definitive Pronouns Rules *

* Verb going to the end from English → Hindi, Marathi, Manipuri - is called Structural Divergences

* very soft changing to soft soft (Reduplication) in Manipuri is called Lexical Divergences

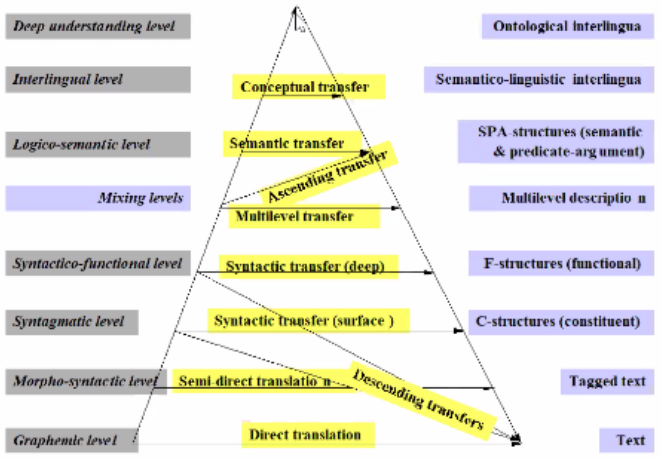
In ML algorithms will pick up this patterns automatically (NN's)

* NN's are data hungry & all Rule Based systems have the problem of false +ve's & false -ve's.

Vauquois Triangle

Kinds of MT Systems

(point of entry from source to the target text)



(Vauquois. 1968)

cvit-mt:pushpak

Representation)

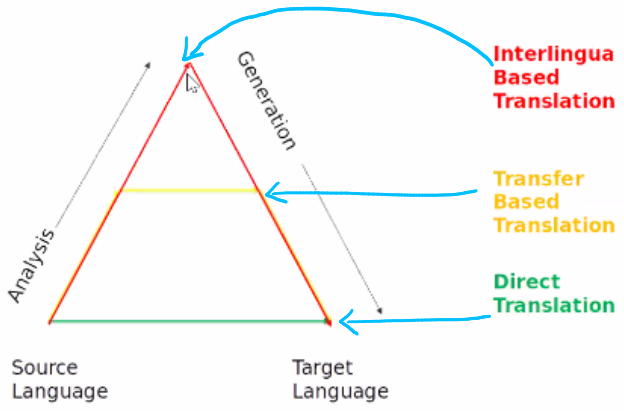
(Disambiguation) Disintegrating the source.

* But be careful not to overkill by complete disambiguation.

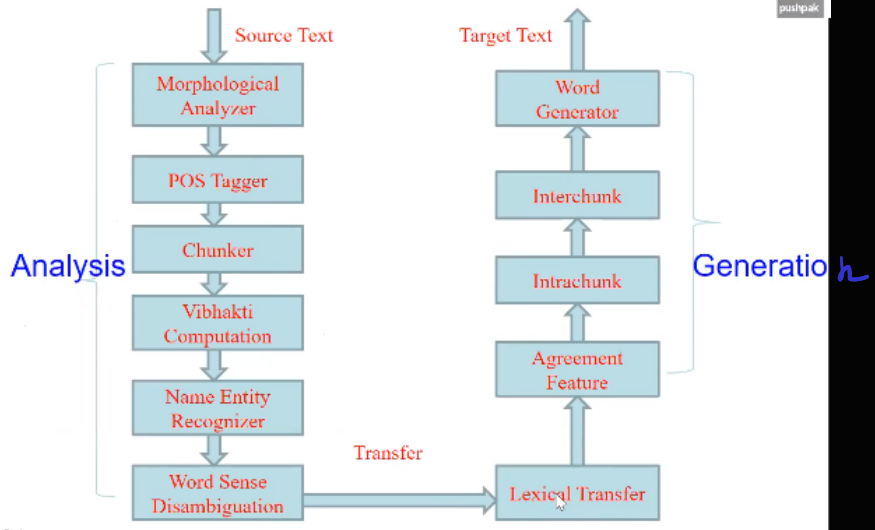
* We can retain ambiguity & still do a successful target sentence generation: {Different languages can work with different levels of ambiguity}

1
0

Simplified Vauquois



Rule based MT (typical architecture)



9aug21

cvit-mt:pushpak

Key Research Areas

- Machine Translation
- Sentiment Analysis
- Information Retrieval
- Lexical Semantics
- Information Extraction
- Cognitive NLP

Linguistics is the eye and computation the body!

Challenges of IL Computing (1/2)

- **Scale and Diversity: 22** major languages in India, written in 13 different scripts, with over 720 dialects
- **Code Mixing** ("kyo ye hesitation?"); **Gerundification** ("gaadi chalaaoing")
- **Absence of basic NLP tools and resources:** ref nlp pipeline
- **Absence of linguistic tradition for many languages**

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga and Ranjiva Munasinghe, *Indic Language Computing*, **CACM**, V 62(11), November 2019.

ILT Challenges (2/2)

- **Script complexity and non-standard input mechanism**: InScript Non-optimal
- **Non-standard transliteration** ("mango"→ 'am", "aam", Am")
- **Non-standard storage**: proprietary fonts
- **Challenging language phenomena**: Compound verbs ("has padaa"), morph stacking ("gharaasamorchyaanii")
- **Resource Scarcity**

} Qwerty keyboard issues with Indian typing

→ Media have their own fonts not following Unicode standards

→ tokens stacked together → word

Indian Language SMT (2014)

	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.

Indo Aryan family

Dravidian

→ scores fell due to Dravidian influence

→ Extreme agglutination

* MT will scarce data is a problem → Challenge (*)

Subword Based MT came into picture after this results (specially with the Dravidian family)

Mitigating the Resource Problem

Three ways (1/2)

(1) Artificially boost the resource

- Subword based NLP
  - Characters, Syllables, Orthographic Syllables, Byte Pair Encoding
    - *Given, "khaa+uMgaa → will+eat" AND "jaa+rahaa_hE → is+going"*
    - *Produce "khaa+rahaa_hE → is+eatin*

→ doesn't appear in training but subword training helps

↓

needs to be a scoring to prevent non linguistic constructs.

Three ways (2/2)

(2) Take help from another language

- Cooperative NLP

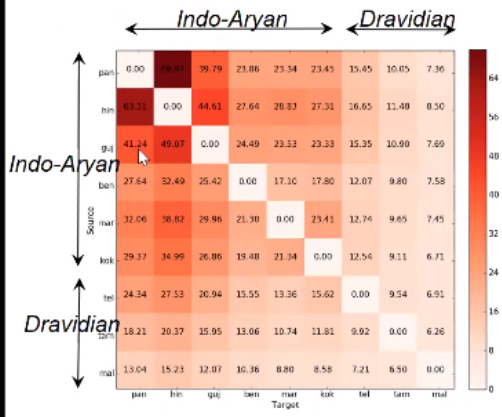
(3) Use "higher level language properties"

e.g., Part of Speech, Sense ID etc.

But there is a pitfall- NLP's "Law of Trade off"

- Trade Off:
  - –*Precision vs. Recall*
  - –Sparsity *vs.* Ambiguity
  - –Information_Injection *vs.* Topic_Drift

Word level translation (BLEU score



Clear Partitioning based on language families

Translation between Indo Aryan languages is easiest

Translation into Dravidian languages is particularly difficult

22

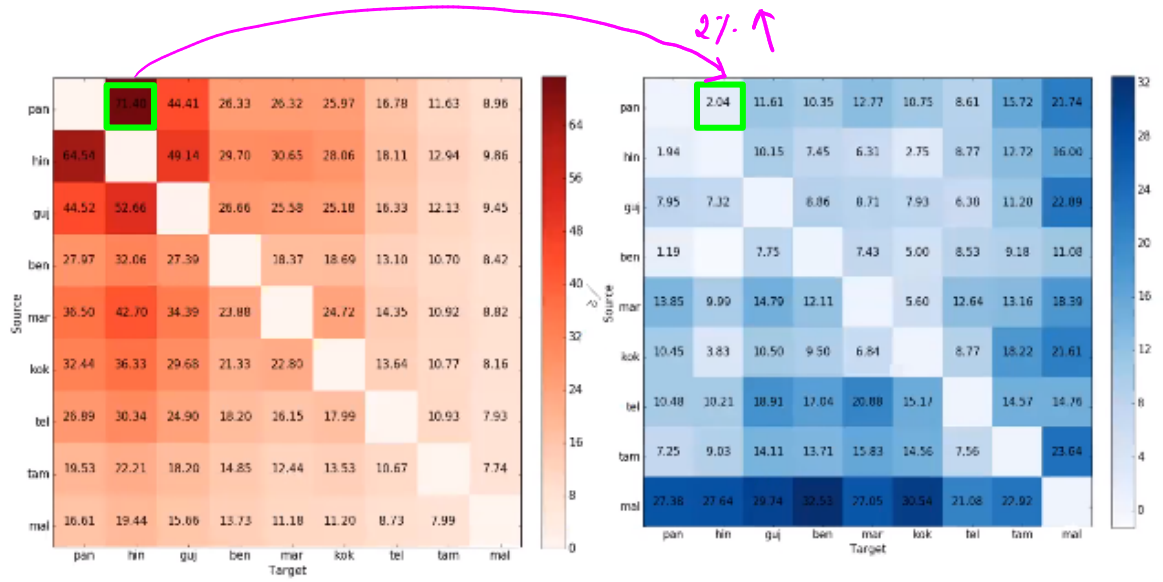
9aug21 cvit-mt:pushpak

Subwords (for "jaauMgaa")

- Characters: "j+aa+u+M+g+aa"
- Morphemes: "jaa"+"uMgaa"
- Syllables: "jaa"+"uM"+"gaa"
- Orthographic syllables: "jaau"+"Mgaa"
- BPE (depends on corpora, statistically frequent patterns): both "jaa" and "uMgaa" are likely

* Breaking the words into syllables is a complex task in itself

Morph level translation



2% ↑

BLEU scores

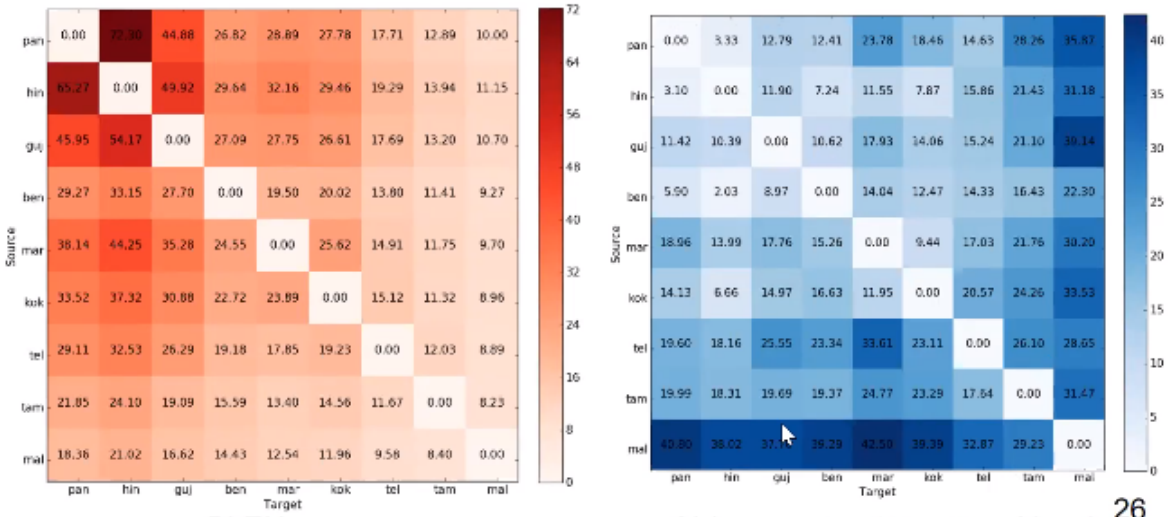
% improvement over word level scores

25

9aug21 cvit-mt:pushpak

Subword Based Translation.
(∵ Subwords can isolate the morphemes

BPE level translation



BLEU scores

% improvement over word level scores

26

9aug21 cvit-mt:pushpak

more ↑ % seen.

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.

Semantic relations+Suffixes→Case Markers+inflections

I *ate* *mangoes*

agent

I {<agt} *ate {eat@past}* *mangoes {<obj}*

I {<agt} *mangoes {<obj.@pl}* *{eat@past}*

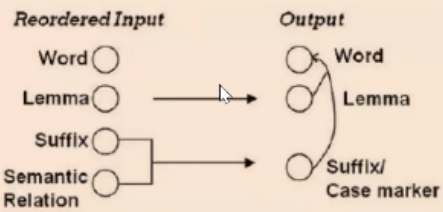
named marker *

Hindi → *mei_ne* *aam* *khaa_yaa*

Rule Operated

* as the verb are in transitive.

Our Factorization based on Koehn and Hoang (2007)



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*))

* Read ↙

Experiment: Corpus Statistics

Very small corpus

	#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

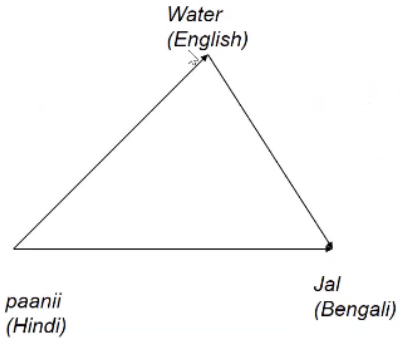
Subjective Evaluation: The impact of reordering and semantic relations

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

Cooperative NLP: Pivot Based MT

Raj Dabre, Fabien Cromiere, Sadao Kurohash and Pushpak Bhattacharyya, *Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages*, **NAACL 2015**, Denver, Colorado, USA, May 31 - June 5, 2015.

Triangulation



L1→bridge→L2 *(Wu and Wang 2009)*

- Resource rich and resource poor language pairs
- Question-1: How about translating through a 'bridge'?
- Question-2: how to choose the bridge?

Mathematical preliminaries

$$e_{best} = \arg\max_{e} p(e|f) = \arg\max_{e} p(f|e) p_{LM}(e)$$

Where $p(f|e)$ is given by:

$$p(\mathbf{f}|\mathbf{e}) = p\left(\bar{f}_1^I \middle| \bar{e}_1^I\right) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) p_w(\bar{f}_i|\bar{e}_i, a)^{\gamma}$$

$$\phi(\bar{f}_i|\bar{e}_i) = \sum_{\bar{p}_i} \phi(\bar{f}_i|\bar{p}_i) \phi(\bar{p}_i|\bar{e}_i)$$

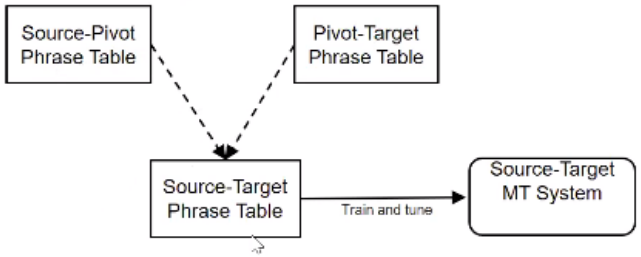
$$p_w(\bar{f}_i|\bar{e}_i, a) = \prod_{l=1}^{n} \frac{1}{|m/(l,m)\in a|} \sum_{\forall(l,m)\in a} w(f_l|e_l)$$

f = target language sentence

e = source language sentence

* Introduction of pivot in scarce resource problem is helpful.
↓
few shot ML

Triangulation approach



- Important to induce language dependent components such as phrase translation probability and lexical weight

Mauritian Creole (MCR) →French (FR) →English (E)

- MCR and FR share vocabulary and structure

Vocabulary matching

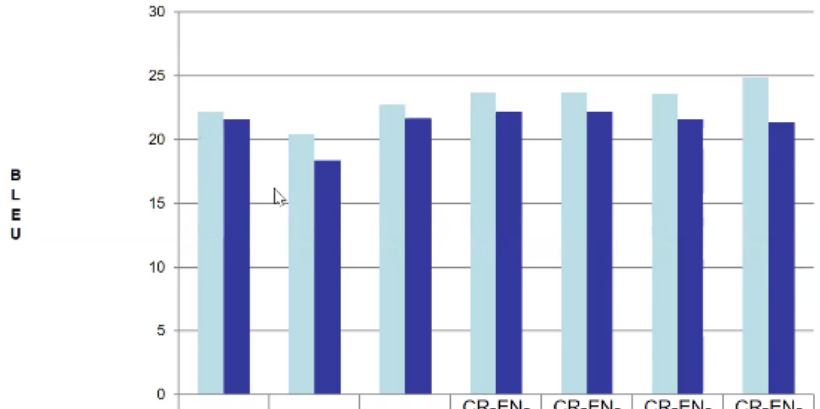
French	Creole	English
avion	Avion	aeroplane
bon	Bon	good
gaz	Gaz	gas
bref	bref	brief
pion	pion	pawn

Source → pivot → target

Experiment on MCR→FR→E

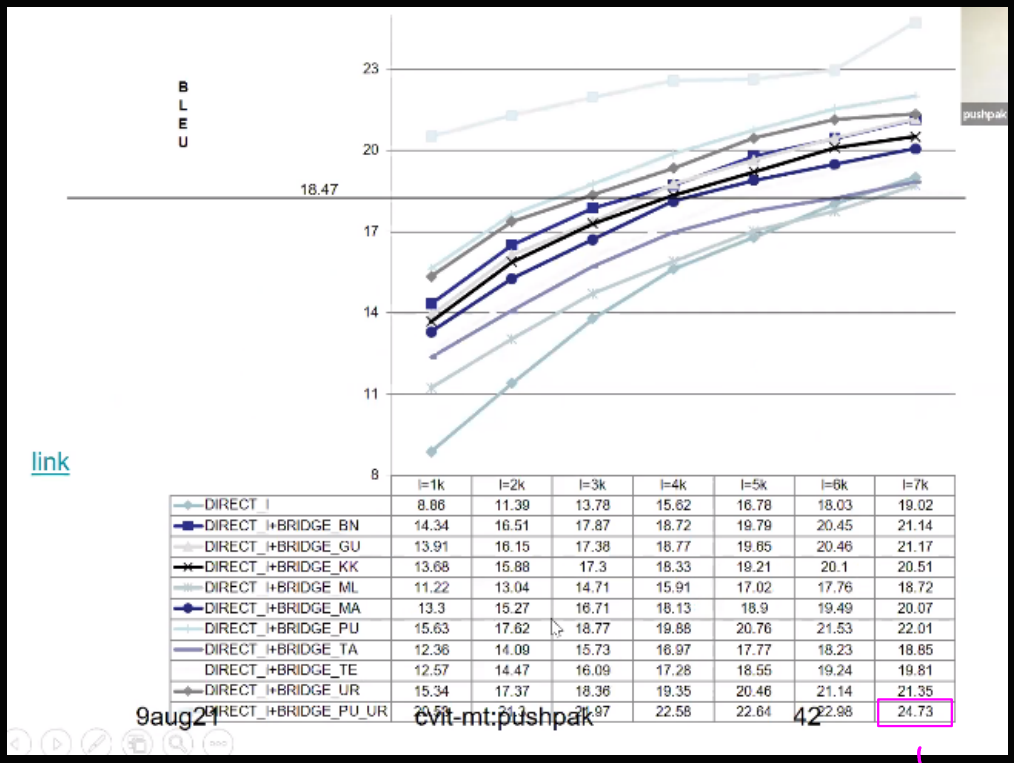
Language pair	#Sentences	#unique words (L1-L2)
En-Fr	2000000	127405- 147812
En-Cr (train + tune)	25010	16294-17389
En-Cr (test)	284 (142 short + 142 long)	1168-1070 + 3562-3326
Fr-Cr	18354	13769-13725

Results



	CR-EN-BASIC	CR-EN-MOD	CR-EN-COMBO	CR-EN-BACKOFF-1	CR-EN-BACKOFF-2	CR-EN-BACKOFF-3	CR-EN-BACKOFF-4
Short-Sentences	22.13	20.44	22.74	23.69	23.69	23.58	24.9
Long Sentences	21.61	18.37	21.66	22.2	22.2	21.59	21.36

Long Sentences showed better results



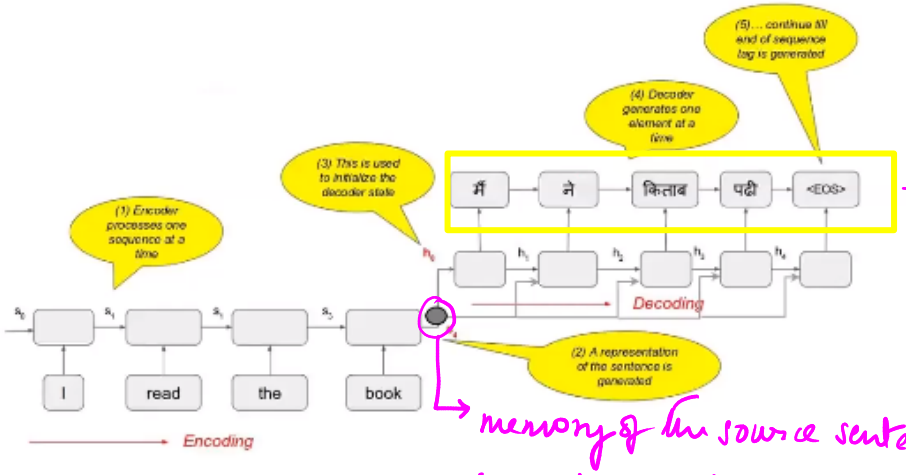
link

	l=1k	l=2k	l=3k	l=4k	l=5k	l=6k	l=7k
DIRECT_I	8.86	11.39	13.78	15.62	16.78	18.03	19.02
DIRECT_I+BRIDGE_BN	14.34	16.51	17.87	18.72	19.79	20.45	21.14
DIRECT_I+BRIDGE_GU	13.91	16.15	17.38	18.77	19.65	20.46	21.17
DIRECT_I+BRIDGE_KK	13.68	15.88	17.3	18.33	19.21	20.1	20.51
DIRECT_I+BRIDGE_ML	11.22	13.04	14.71	15.91	17.02	17.76	18.72
DIRECT_I+BRIDGE_MA	13.3	15.27	16.71	18.13	18.9	19.49	20.07
DIRECT_I+BRIDGE_PU	15.63	17.62	18.77	19.88	20.76	21.53	22.01
DIRECT_I+BRIDGE_TA	12.36	14.09	15.73	16.97	17.77	18.23	18.85
DIRECT_I+BRIDGE_TE	12.57	14.47	16.09	17.28	18.55	19.24	19.81
DIRECT_I+BRIDGE_UR	15.34	17.37	18.36	19.35	20.46	21.14	21.35
DIRECT_I+BRIDGE_PU_UR	15.53	17.3	18.97	22.58	22.64	22.98	24.73

Punjabi & Urdu as pivots → good score

Neural Machine Translation

Encoder-Decoder model



Softmax
Prob. highest

memory of the source sentence is compromised.

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

↓

Attention Networks (Transformers) came into picture.

Some representative accuracy figure for Indian Language NMT (highly constrained domain)

Language pair	BLEU score
Hi - Mr	31.25
Hi - Pa	63.38
Pa - Hi	68.31
Hi - Gu	49.98
Gu - Hi	**53.22** (↑ from 53.09 from SMT)

Unsupervised Neural Machine Translation

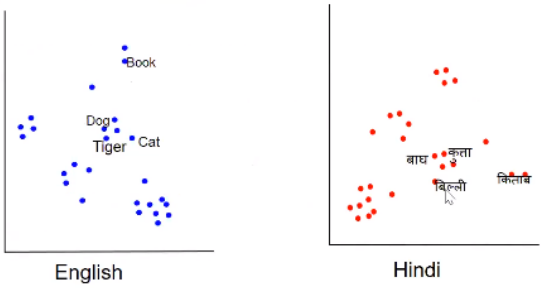
Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal and Pushpak Bhattacharyya. Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders. 57th Annual Meeting of the Association for Computational Linguistics (**ACL 2019**), Florence, Italy. 2019

Background: Unsupervised NMT (Artetxe et al., 2018)

1. Consists of one shared Encoder and multiple Decoders
2. Fixed cross-lingual embedding at encoder side
3. Denoising Auto-encoding
4. Back-translation

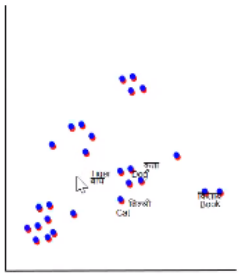
Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. ICLR

Background: Geometric Structure of Words (use iso-metricity)



After Cross-lingual Mapping

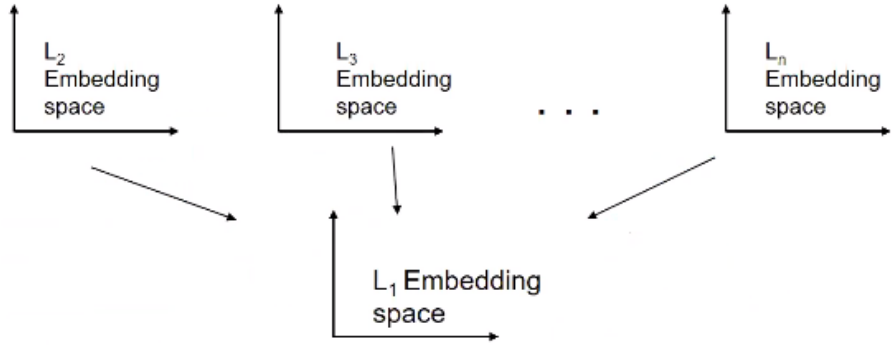
This involves strong assumption that embedding spaces across languages are isomorphic, which is not true specifically for distance languages (Søgaard et al. 2018). However, without this assumption unsupervised NMT is not possible.



Søgaard, Anders, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. ACL

$L_1 \rightarrow$ English.
$L_2, L_3, L_4 \rightarrow$ Non English.

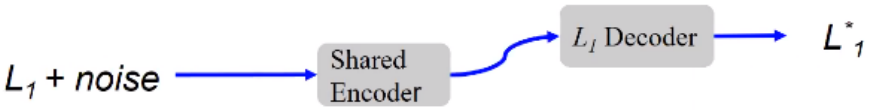
Multilingual Embedding



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

*

Denoising Auto-encoding for L1



Loss between L^*_1 and L_1

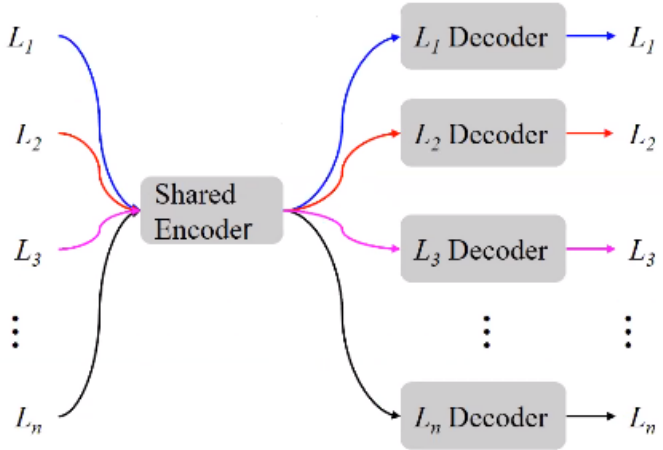
Noise is introduced through swapping/deletion of words. For example,

Original: **An investment in knowledge pays the best interest .**

After noising: **investment An in knowledge pays best the interest .**

Model tries to predict original from noisy sentence. We do it for sentences in all four languages.

Denoising and Back-Translation



* Takes cross path via the shared Encoder. *

Results

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

Table 1: BLEU scores on *newstest2013* and *newstest2014*. ▲ shows improvements over bilingual models. Spanish (Es) is not part of the *newstest2014* test set. **Base**: Baseline. **Multi**: Multingual

Sample Outputs (French→ English)

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

Sample Outputs (Spanish→ English)

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

Sample Outputs (German→ English)

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

Results (Zero-shot Translation in Unsupervised NMT)

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

Table 2: BLEU scores of translation between non-English languages on *newstest2013*. Consider rows are source and columns are target. The network is not trained for these language pairs and still it is possible to translate between these pairs by using the shared encoder and language specific decoders.

Sample Outputs (Unseen pairs)

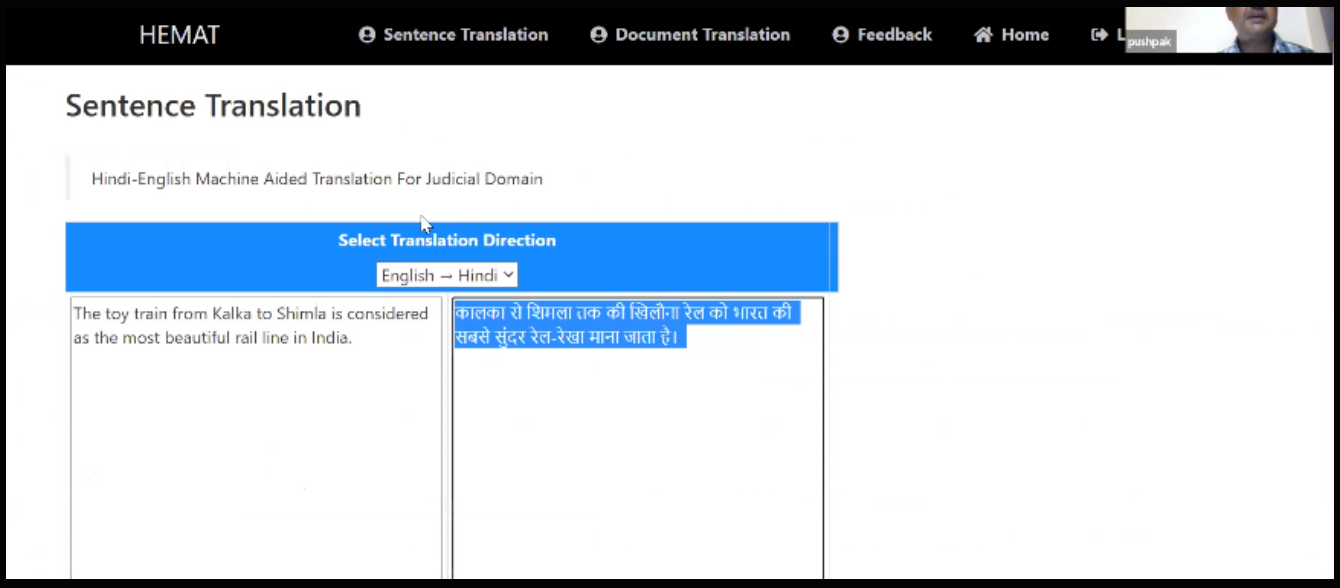
Source	Reference	Multilingual
	French→Spanish	
Les dirigeants républicains justifièrent leur politique par la nécessité de lutter contre la fraude électorale.	Los dirigentes republicanos justificaron su política por la necesidad de luchar contra el fraude electoral.	Los dirigentes republicanos <OOV> su política por la necesidad de luchar contra la fraude electoral.
	French→German	
Chacun sait que son livre fait partie de cet édifice.	Jeder weiß , dass sein Buch Teil dieses Gebäudes ist.	Jeder weiß , dass sein Buch Teil seines Gebäudes machte.
	German→Spanish	
Seine Zahlen auf Ebene der internationalen Turniere sind beeindruckend.	Sus números a nivel de torneos internacionales son impresionantes.	Sus cifras sobre el nivel de torneos internacionales son impresionantes.
	German→French	
Diese Einschränkungen sind nicht ohne Folgen.	Ces restrictions ne sont pas sans conséquence.	Ces restrictions ne sont pas sans conséquences.
	Spanish→German	
Tomemos por caso la elección directa del presidente , que ha sido un logro de la presión pública.	Nehmen Sie nur einmal die direkte Wahl des Präsidenten, die ein Verdienst des öffentlichen Drucks war.	Nehmen Sie über die direkte Wahl des Präsidenten, hat dies ein Erfolg ein der öffentlichen Druck.
	Spanish→French	
Las inversiones en la materia superan los 1.5 billones de dólares.	Les investissements dans ce domaine dépassent les 1,5 milliards de dollars.	Les investissements dans la matière dépassent les 1,5 milliards de dollars.

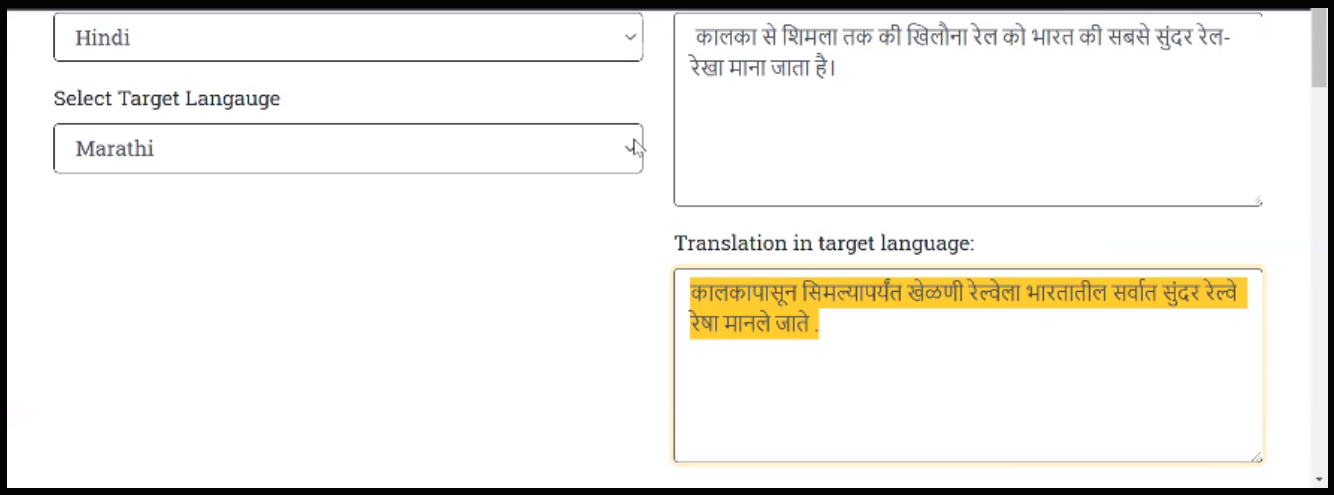
Unsupervised NMT with filtering on Back Translation

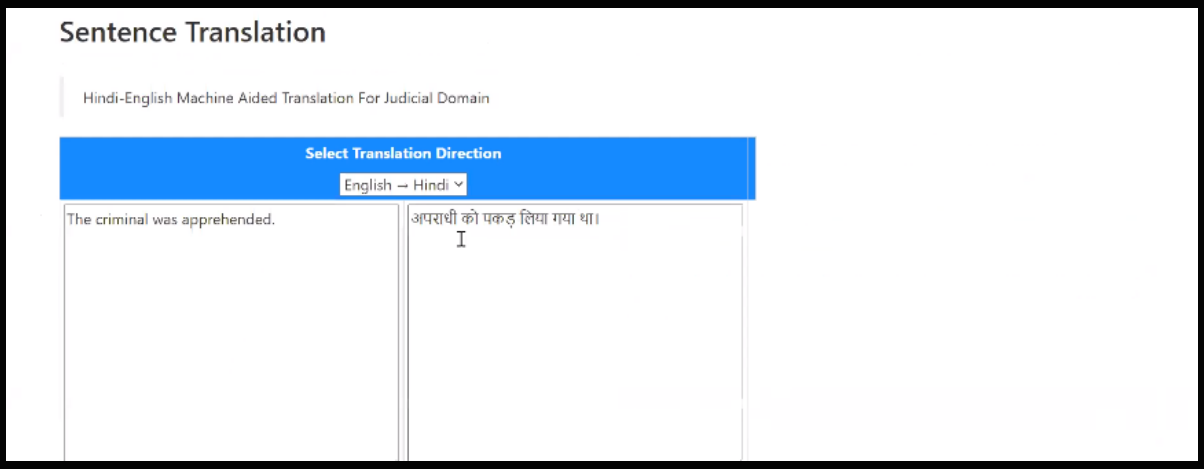
Jyotsana Khatri and Pushpak Bhattacharyya, *Filtering Back-Translated Data in Unsupervised Neural Machine Translation* , 28th Int'l Conf on Computational Linguistics (**COLING20**), Online Conference, December 8-13, 2020.

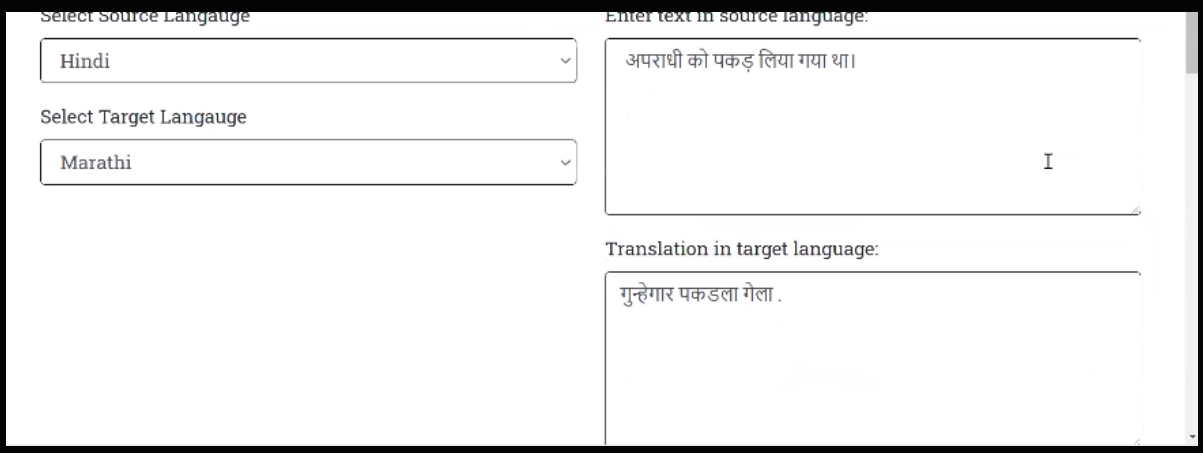
Central Idea

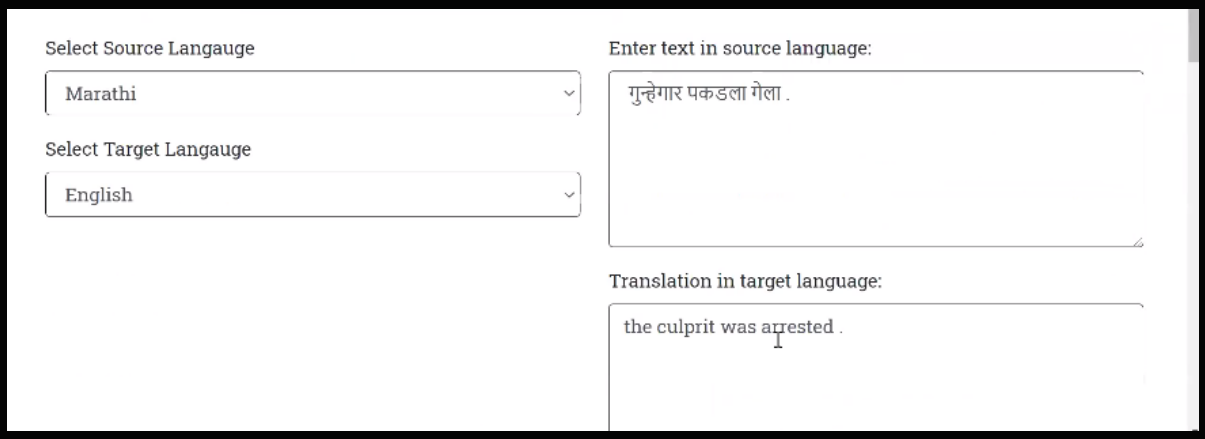
- Filter back-translated data by giving more weight to good pseudo parallel sentence pairs.
- Quality of pseudo parallel sentence pairs measured using round trip BLEU







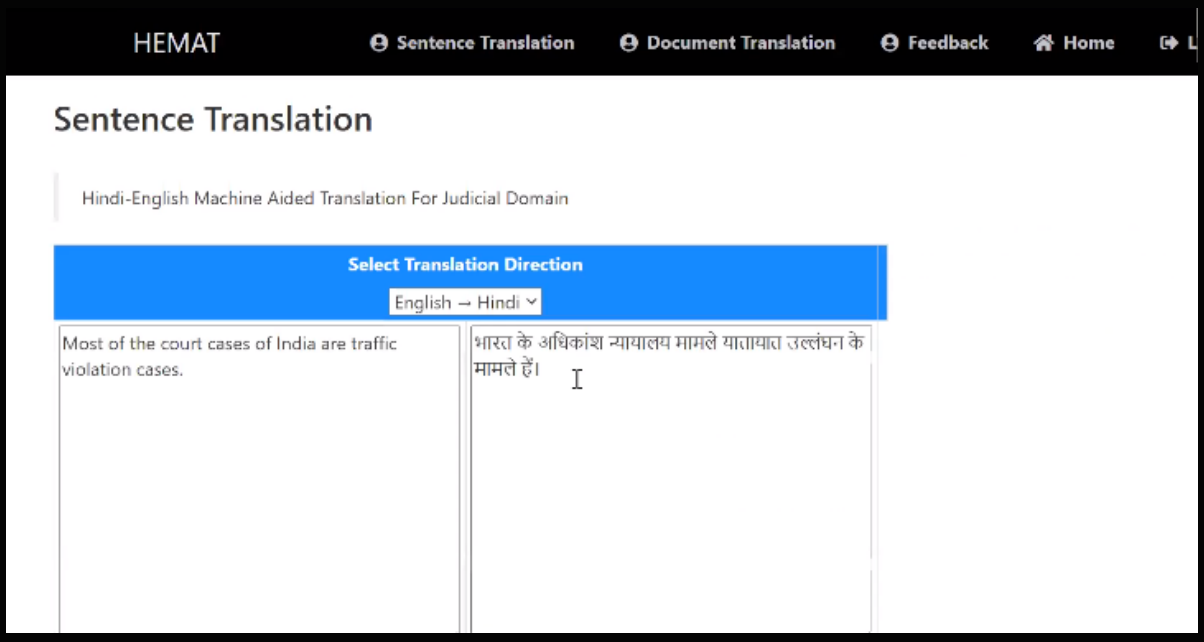


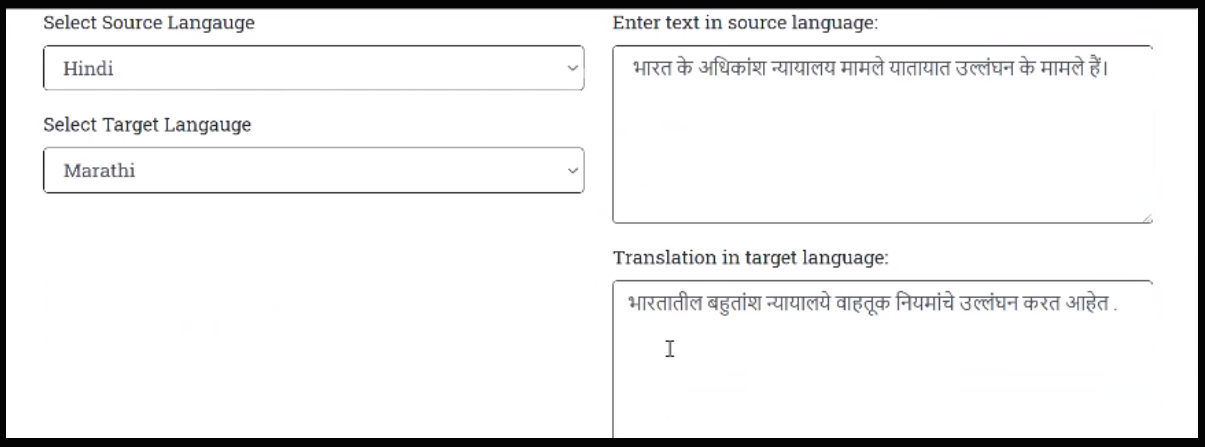


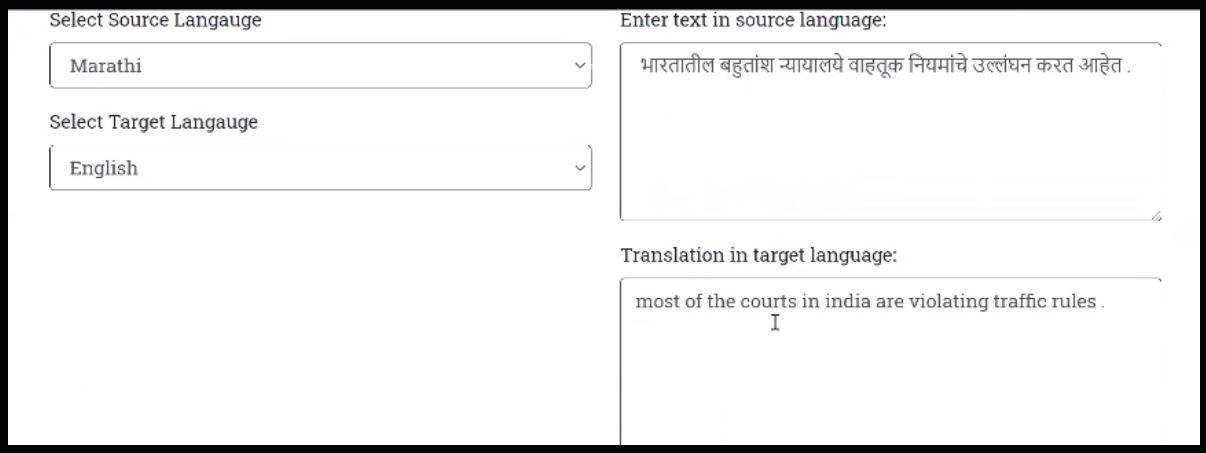
* Meaning is serviced but this method of evaluation by completing a circle for a pair of sentence is a stringent form of evaluation. {or else errors can accumulate to give bad target o/p}

Another Example; when o/p is disturbed {post editing is required}

* Work Error Rate
* Cognitive Based Load
* Eye Tracking based Cognitive Load.

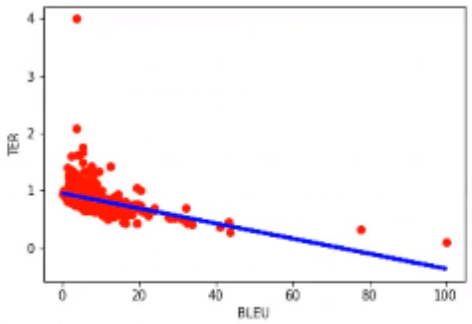






* Postediting: Most of the court cases in India are for violating traffic rules

Bible



Equation of Line : y = -0.013x + 0.955

Slope(m) = -0.013

Intercept(c) = 0.955

Lang-pair	# sentences in source language	# sentences in target language	Training dataset
en-fr	5M	5M	WMT News Crawl
en-de	5M	5M	WMT News Crawl
en-ro	5M	2.28M	WMT News Crawl

Table 1:Details of training data

Lang-pair	Validation data	Test data
en-fr	Newstest 2013	Newstest 2014
en-de	Newstest 2013	Newstest 2016
en-ro	Newsdev 2016	Newstest 2016

Table 2:Details of test data

Results

Method	en-fr	fr-en	en-ro	ro-en	en-de	de-en
Song et al., 2019	26.59	25.42	25.53	24.8	17.62	24.78
Song et al., 2019 + Filtering	26.37	25.5	25.29	24.64	17.51	24.87
Lample and Conneau, 2019	27.95	27.02	26.26	25.8	18.26	25.22
Lample and Conneau, 2019 + Filtering	28.4*	27.69*	26.96*	26.24*	17.35	25.91*

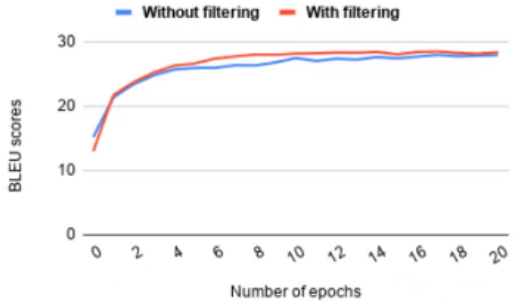


Figure 1:Progress with iterations: BLEU scores for each epoch for en-fr

Neural Indian Language MT

Comprehensive ILNMT, including some points from the bahubhashak project

Shubham Dewangan, Shreya Alva, Nitish Joshi and Pushpak Bhattacharyya, *Experience of Neural Machine Translation between Indian Languages*, Journal of Machine Translation (**JMT**), Springer, Accepted. DoI

Recent Publications

- Shubham Dewangan, Shreya Alva, Nitish Joshi and Pushpak Bhattacharyya, *Experience of neural machine translation between Indian languages*, **Machine Translation volume 35, 2021.**
- Akash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal, Pushpak Bhattacharya, *Neural Machine Translation in Low-Resource Setting: a Case Study in English-Marathi Pair*, **MT Summit 2021**.
- Tamali Banerjee, Rudra V Murthy, Pushpak Bhattacharya, *Crosslingual Embeddings are Essential in UNMT for distant languages: An English to IndoAryan Case Study*, **MT Summit 2021**.
- Kamal Gupta, Dhanvanth Boppana, Rejwanul Haque, Asif Ekbal, Pushpak Bhattacharyya, *Investigating Active Learning in Interactive Neural Machine Translation*, **MT Summit 2021**.
- Tamali Banerjee, Rudra V Murthy, Pushpak Bhattacharya, Scrambled Translation Problem: A Problem of Denoising UNMT, **MT Summit 2021**.

*

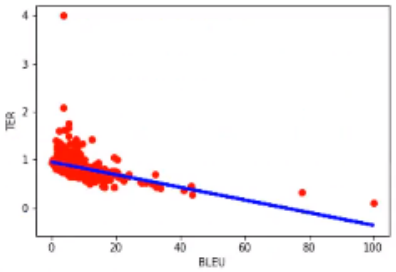
TER → Transmission Error Rate

When

TER = 1, BLEU = 0

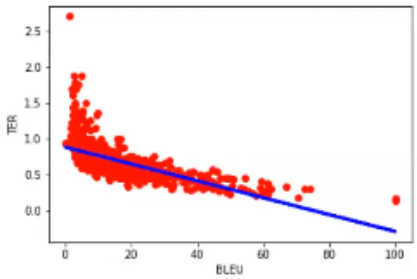
= 0, = 1

Bible



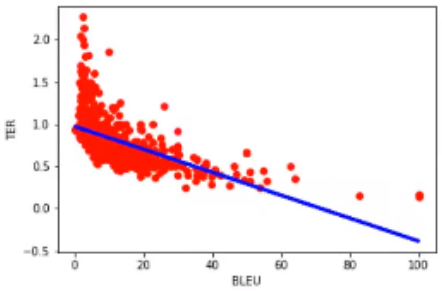
Equation of Line : y = -0.013x + 0.955
Slope(m) = -0.013
Intercept(c) = 0.955

ILCI



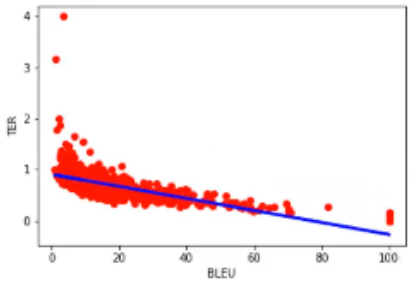
Equation of Line : y = -0.011x + 0.880
Slope(m) = -0.011
Intercept(c) = 0.880

PIB



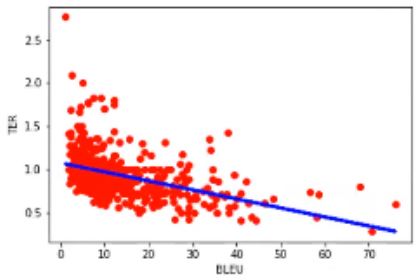
Equation of Line : y = -0.013x + 0.974
Slope(m) = -0.013
Intercept(c) = 0.974

PMI



Equation of Line : y = -0.011x + 0.912
Slope(m) = -0.011
Intercept(c) = 0.912

Tourism



Equation of Line : y = -0.010x + 1.075
Slope(m) = -0.010
Intercept(c) = 1.075

Summary

- MT Paradigms
- Data Driven MT: SMT and NMT
- Tricks of Resource Mitigation
- Unsupervised NMT
- Experience of IL-NMT

Summary on resource mitigation tricks

- Several techniques explored and demonstrated their efficacy.
  - Phrase Table Injection, has great potential to boost BLEU scores, particularly when Dravidian languages are involved.
  - Harnessing monolingual data with back translation, forward translation is advantageous.
  - Enhancements like morph and word feature injection

Conclusions and Future Work

1. Data needed for quick system development, robustness
2. Linguistic insight needed for data curation, annotation and insightful error analysis
3. NMT ensured fluency- syntactic goodness
4. No way out other than various tricks for amelioarting data hunger

Future lies in:

Cross Lingual, Multilingual Word embedding

Final Message

"NLP is a task in Trade Off"

e.g., Not too much of subwords or cooperation

(beware of *'ambiguity insertion'*),

not too little

(beware of 'sparsity') !!