CS626: Speech, NLP and Web

Machine Translation- Language Divergence, Evaluation, Bridge Problem

Pushpak Bhattacharyya

Computer Science and Engineering Department

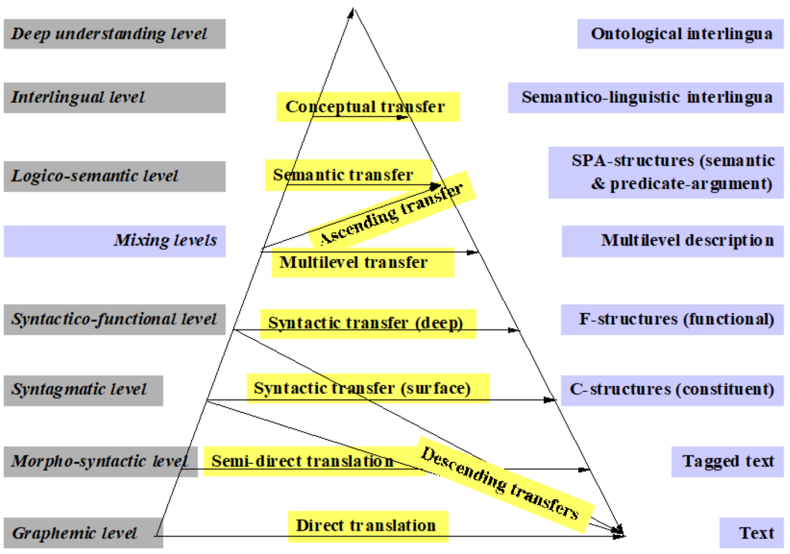
IIT Bombay

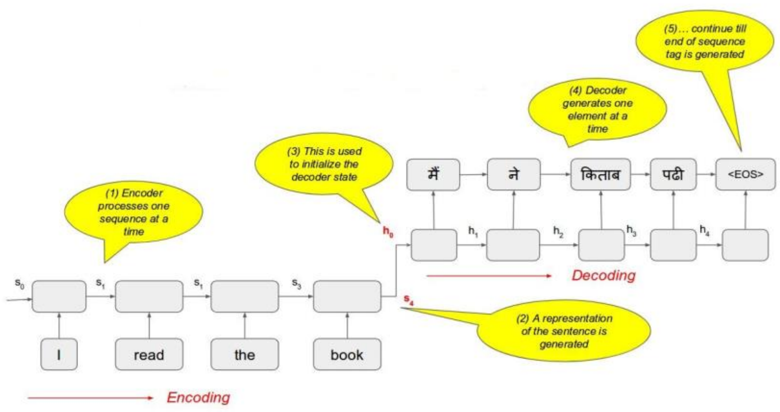
Week 9 of 30th September, 2024

1-slide recap of week of 2nd Sep

- Machine Translation: Definition, Paradigms
- Main Challenge: Language Divergence
- Vauquois Triangle as an abstraction of paradigms of MT
- A-T-G framework: Analysis Transfer Generation
- Encode Decoder Framework: basis of neural MT
- Data Driven MT- noisy channel model-

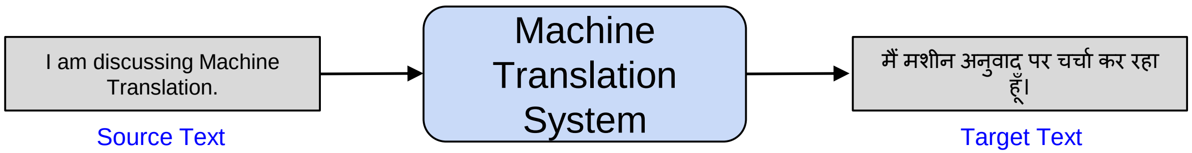
$$\bar{e} = \arg\max_{e} P(e|f)$$





Machine Translation

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



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

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

Paradigms of Machine Translation

← Human effort

Data hungry →

RBMT → EBMT → SMT → NMT

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

Rules handmade by human expert → Rules by learning from data

Main Challenge of MT: Language Divergence

Kinds of MT Systems

(point of entry from source to the target text)

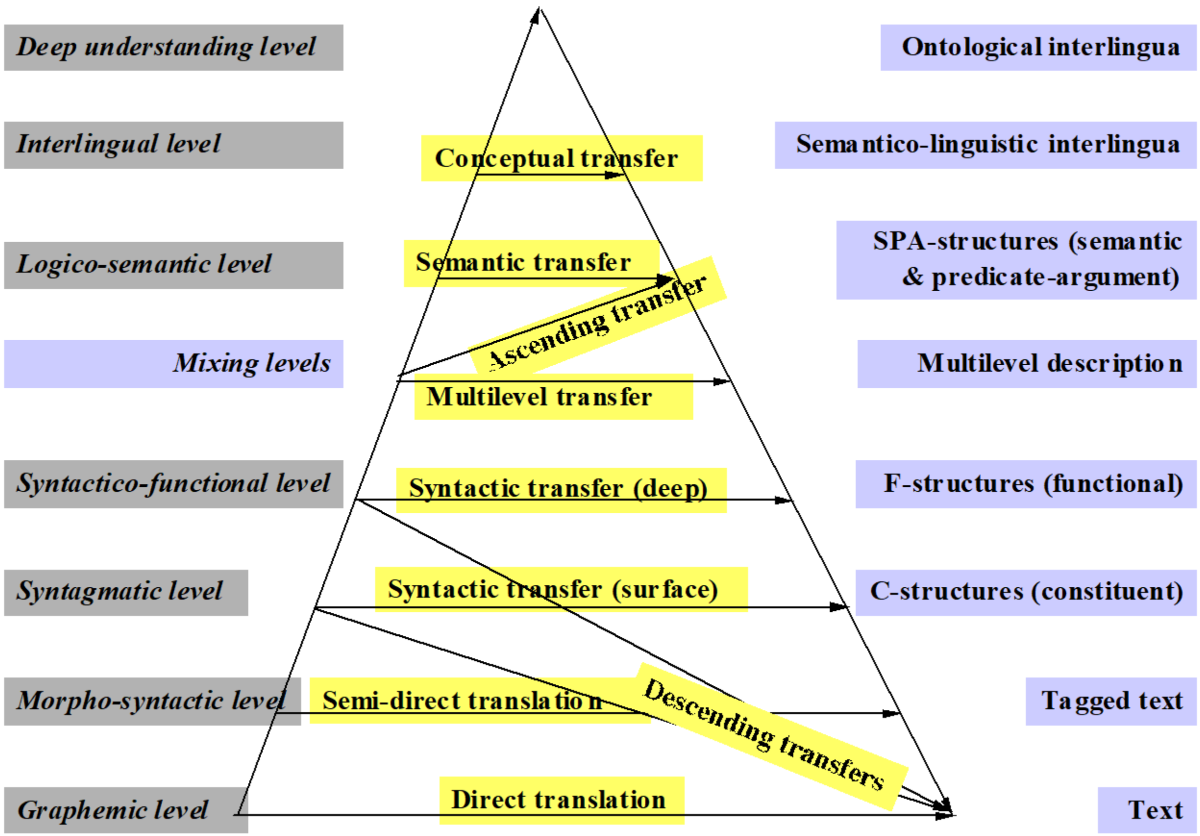
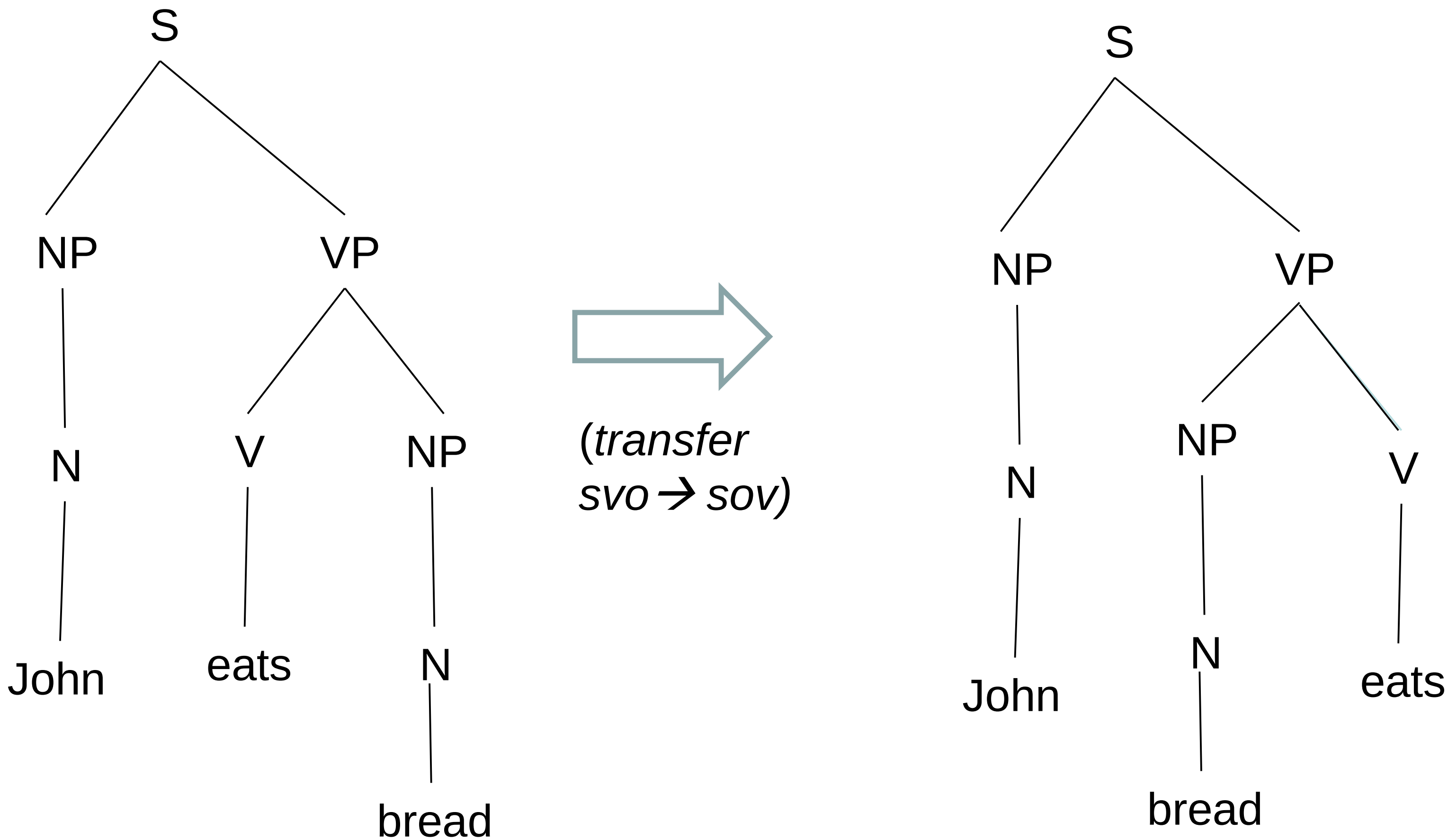


Illustration of transfer SVO→SOV



Understanding the Analysis-Transfer-Generation over Vauquois triangle (1/4)

H1.1: सरकार_ने चुनावो_के_बाद मुंबई में करों_के_माध्यम_से अपने राजस्व_को बढ़ाया |

T1.1: Sarkaar ne chunaawo ke baad Mumbai me karoM ke maadhyam se apne raajaswa ko badhaayaa

G1.1: Government_(ergative) elections_after Mumbai_in taxes_through its revenue_(accusative) increased

E1.1: The Government increased its revenue after the elections through taxes in Mumbai

Understanding the Analysis-Transfer-Generation over Vauquois triangle (2/4)

Entity	**English**	**Hindi**
Subject	The Government	सरकार (sarkaar)
Verb	Increased	बढ़ाया (badhaayaa)
Object	Its revenue	अपने राजस्व (apne raajaswa)

Understanding the Analysis-Transfer-Generation over Vauquois triangle (3/4)

Adjunct	**English**	**Hindi**
Instrumental	Through taxes in Mumbai	मुंबई_में करों_के_माध्यम_से (mumbai me karo ke maadhyam se)
Temporal	After the elections	बढ़ाया (badhaayaa)

Understanding the Analysis-Transfer-Generation over Vauquois triangle (3/4)

The Government increased its revenue

P0 P1 P2 P3

E1.2: after the elections, the Government increased its revenue through taxes in Mumbai

E1.3: the Government increased its revenue through taxes in Mumbai after the elections

More flexibility in Hindi generation

Sarkaar_ne badhaayaa

P_0 (the govt) P_1 (increased) P_2

H1.2: चुनावो_के_बाद सरकार_ने मुंबई_में करों_के_माध्यम_से अपने राजस्व_को बढ़ाया |

T1.2: elections_after government_(erg) Mumbai_in taxes_through its revenue increased.

H1.3: चुनावो_के_बाद मुंबई_में करों_के_माध्यम_से सरकार_ने अपने राजस्व_को बढ़ाया |

T1.3: elections_after Mumbai_in taxes_through government_(erg) its revenue increased.

H1.4: चुनावो_के_बाद मुंबई_में करों_के_माध्यम_से अपने राजस्व_को सरकार_ने बढ़ाया |

T1.4: elections_after Mumbai_in taxes_through its revenue government_(erg) increased.

H1.5: मुंबई_में करों_के_माध्यम_से चुनावो_के_बाद सरकार_ने अपने राजस्व_को बढ़ाया |

T1.5: Mumbai_in taxes_through elections_after government_(erg) its revenue increased.

What is the main challenge of MT?
Language Divergence

Syntactic Divergence

Constituent Order divergence
Adjunction Divergence
Preposition-Stranding divergence
Movement divergence
Null Subject Divergence
Dative Divergence
Pleonastic Divergence

Constituent Order Divergence

E24. <u>Jim</u> <u>is playing</u> <u>tennis</u>.

S V O

H24. <u>जीम</u> <u>टेनिस</u> <u>खेल रहा है</u>।

<u>jeem</u> <u>Tenis</u> <u>khel rahaa hai</u>

<u>Jim</u> <u>tennis</u> <u>playing-is</u>

S O V

E25. <u>He</u> <u>saw</u> <u>a girl</u> <u>whose eyes were blue.</u>

S V O Q

H25. <u>उस ने</u> <u>एक</u> <u>लड़की को</u> <u>देखा</u> <u>जिसकी ऑखें नीली थी</u>।

<u>us ne</u> <u>ek ladakee ko</u> <u>dekhaa</u> <u>jisakee aankhen neelee thee</u>

<u>He</u> <u>a girl-to</u> <u>saw</u> <u>whose eyes blue were</u>

S O V Q

Adjunction Divergence

E26. *the [living in Delhi] boy

H26. [दिल्ली में रहनेवाला] लड़का

[dillee mein rahanevaalaa] ladakaa

[Delhi-in living] boy

H27. (A) राम ने [मोहन को पसंद आनेवाला] तोहफ़ा भेजा।

raam ne [mohan ko pasand aanevaalaa] tohafaa bhejaa

Ram [Mohan-to like come-ing] gift sent

(B) राम ने वह तोहफ़ा भेजा जो मोहन को पसंद आया।

raam ne vah tohafaa bhejaa jo mohan ko pasand aayaa

Ram that gift sent that Mohan-to like came

(C) राम ने वह तोहफ़ा भेजा जो मोहन को पसंद है।

raam ne vah tohafaa bheejaa jo mohan ko pasand hai

Ram that gift sent that Mohan-to like is

E27. Ram sent the gift that mohan likes.

PP Adjunction Divergence

E28. He called me [to his house.]

*He called [to his house] me.

H28. (A) उसने मुझे [अपने घर] बुलाया।

usne mujhe [apne ghar] bulaayaa

he to-me his house called

(B) उसने [अपने घर] मुझे बुलाया।

usne [apne ghar] mujhe bulaayaa

he his house to-me called

Preposition Stranding Divergence

E29. Which shop did John go to?

H29. *<u>किस</u> <u>दुकान</u> <u>ज्होन</u> <u>गया</u> <u>में</u> ?

<u>kis</u> <u>dukaan</u> <u>john</u> <u>gayaa</u> <u>mein</u>

Null Subject Divergence

E30. Long ago, there was a king.

H30. बहुत पहले एक राजा था।

bahut pahale ek raajaa thaa

long ago one king was

H31. जा रहा हूँ।

jaa rahaa hun

going-am

E31. *am going.

Pleonastic Divergence

E32. It is raining.

H32. बारिश हो रही है।

Lexical Semantic Divergence

- Conflational divergence
- Structural divergence
- Categorial divergence
- Head swapping divergence
- Lexical divergence

Conflational Divergence

E33. Jim stabbed John.

H33. जीम ने ज्होन को छूरे से मारा ।

jeem ne john ko chhoore-se maaraa

Jim John-to knife-with hit

Structural Divergence

E34. Jim entered the house.

H34. जीम घर में प्रवेश किया ।

jeem ghar mein pravesha kiyaa

Jim house-into entry did

Categorial Divergence

E35. They are competing.

H35. वह मुकाबला कर रहे है।

vaha muqaabalaa kar rahe hai

They competition doing-are

Categorial Divergence: demotional

E36. It suffices.

H36. <u>यह</u> <u>काफी है</u> ।

<u>yaha</u> <u>kaafee hai</u>

<u>It</u> <u>sufficient-is</u>

Categorial Divergence-Promotional

E37. The play is on.

H37. <u>खेल</u> <u>चल रहा है</u> ।

Lexical Divergence

H38.	ज्होन	जबरजस्ती	घर में	घुस गया।
	john	jabarjasti	ghar mein	ghus gayaa
	John	forcefully	house-in	enter-go

E38. John broke into the house.

MT evaluation

Evaluation in MT

- Operational evaluation
  - “Is MT system A operationally better than MT system B? Does MT system A cost less?”
- Typological evaluation
  - “Have you ensured which linguistic phenomena the MT system covers?”
- Declarative evaluation
  - “How does quality of output of system A fare with respect to that of B?”

Adequacy (also called comprehensibility, fidelity, faithfulness) and Fluency

- Assign scores to specific qualities of output
  - Fluency: How good the output is as a well-formed target language entity
  - Adequacy: How good the output is in terms of preserving content of the source text

Form Content Dichotomy

- Ancient philosophical concept
- Consider a pot of milk: milk has the form of pot
- Pot has the content as milk.
- Adequacy refers to content, fluency refers to form

Adequacy and Fluency cntd.

For example, I am attending a lecture

मैं एक व्याख्यान बैठा हूँ

Main ek vyaakhyan baitha hoon

I a lecture sit (Present-first person)

I sit a lecture : Adequate but not fluent

मैं व्याख्यान हूँ

Main vyakhyan hoon

I lecture am

I am lecture: fluent but not adequate.

ADEQUACY AND FLUENCY SCALE

Adequacy and Fluency are measured in the scale of 1 to 5.

1: BAD !

2: MEDIOCRE !

3: GOOD !

4: VERY GOOD !

5: EXCELLENT !

What are human evaluators most sensitive to?

Native speakers are particularly keen on the correct usage of morphological variations and function words in the language.

e.g. “Rahul ka behen” instead of “Rahul ki behen” would be critically penalized.

Similarly, “Mary kitab padta hai” rather than “Mary kitab padti hai” would get a much lower score.

BLEU

Used in any kind of natural language generation situation: QA, Summarization, MT, Paraphrasing and so on

Foundational Point

- Human evaluation is the ultimate yardstick
- Any automatic evaluation MUST correlate well with human evaluation
- BLEU for last 20 years has satisfied reasonably this requirement
- Except in case of high morphological complexity, in which case we have to use subword based BLEU

Allied point: IAA

- Human evaluation is the skyline
- But human evaluation is subjective
- We must have multiple evaluators and compute inter-annotator agreement

How is translation performance measured?

The closer a machine translation is to a professional human translation, the better it is.

- A corpus of good quality human reference translations

- A numerical “translation closeness” metric

Suggested Papers

K. Papineni, S. Roukos, T. Ward, and W. Zhu. *Bleu: a method for automatic evaluation of machine translation,* ACL 2002.

Chris Callison-Burch, Miles Osborne, Phillipp Koehn, *Re-evaluating the role of Bleu in Machine Translation Research, European ACL (EACL) 2006, 2006.*

R. Ananthakrishnan, Pushpak Bhattacharyya, M. Sasikumar and Ritesh M. Shah, *Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU*, **ICON 2007**, Hyderabad, India, Jan, 2007.

Cntd.

Preliminaries

- **Candidate Translation(s):** Translation returned by an MT system
- **Reference Translation(s):** ‘Perfect’ translation by humans

Goal of BLEU: To correlate with human judgment

Formulating BLEU (Step 1): Precision

I had lunch now.

Reference 1: मैने अभी खाना खाया

maine abhi khana khaya

I now food ate

I ate food now.

Reference 2 : मैने अभी भोजन किया

maine abhi bhojan kiyaa

I now meal did

I did meal now

Candidate 1: **मैने** अब **खाना खाया**

maine ab khana khaya

I now food ate

I ate food now

matching unigrams: 3,

matching bigrams: 1

Unigram precision: Candidate 1: 3/4 = 0.75, Similarly, bigram precision: Candidate 1: 0.33

Formulating BLEU (Step 1): Precision

I had lunch now.

Reference 1: मैने अभी खाना खाया

maine abhi khana khaya

I now food ate

I ate food now.

Reference 2 : मैने अभी भोजन किया

maine abhi bhojan kiyaa

I now meal did

I did meal now

Candidate 2: **मैने अभी** लंच एट

maine abhi lunch ate matching unigrams: 2,

I now lunch ate

I ate lunch (OOV) now(OOV) matching bigrams: 1

Unigram precision: Candidate 2: 2/4 = 0.5

Similarly, bigram precision: Candidate 2 = 0.33

Precision: Not good enough

Reference: *aapkii badii meharbaanii hogii*

I will be very thankful to you

Candidate 1: *aap badii meharbaanii hogii*

matching unigram: 3

Candidate 2: ***aapkii aapkii aapkii meharbaanii***

matching unigrams: 4

Unigram precision: Candidate 1: 3/4 = 0.75,

Candidate 2: 4/4 = 1

Formulating BLEU: Modified Precision

$Count_{clip}$(n-gram) = min (count, max_ref_count)

Reference: *aapkii badii meharbaanii hogii*

Candidate 2: : ***aapkii aapkii aapkii meharbaanii***

Matching unigrams:

aapkii : min(3, 1) = 1

meharbaaniii: min (1, 1) = 1

Modified unigram precision: 2/4 = 0.5

Modified n-gram precision

For entire test corpus, for a given n

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

n-gram: Matching n-grams in C

n-gram’: All n-grams in C

Precision computation- example

English: I had lunch now.

Reference: मैने अभी खाना खाया

maine abhi khana khaya

Candidate 1: **मैने** अब **खाना खाया**

maine ab khana khaya

matching unigrams: 3

matching bigrams: 1

P1=3/4=0.75

P2=1/3=0.33

Candidate 2: **मैने अभी** लंच एट

maine abhi lunch ate

matching unigrams: 2

matching bigrams: 1

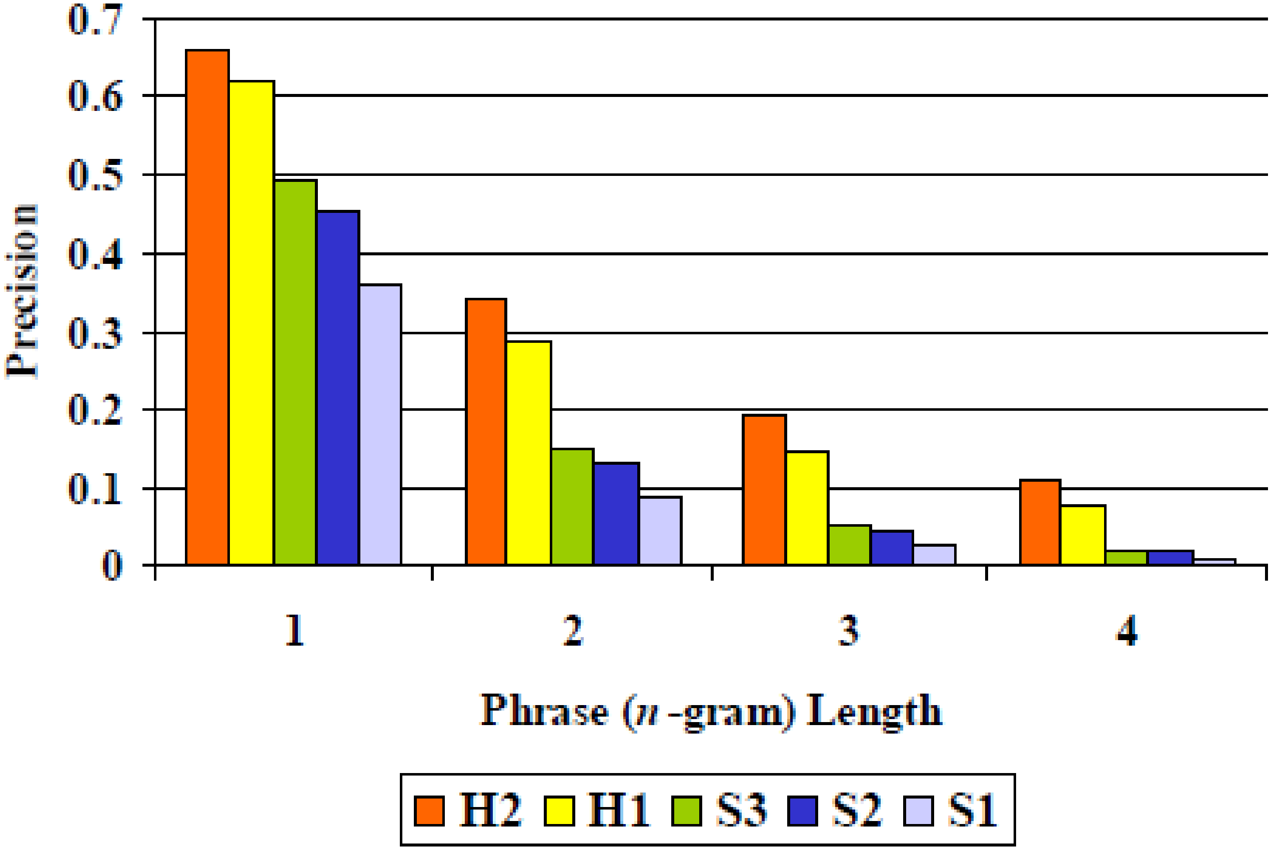
P1=2/4=0.5

P2=1/3=0.33

Comparing HT and MT Precision (1/2)

- From the original BLEU paper (Papineni et al. 2002)
- 127 source sentences were translated by two human translators and three MT systems
- Translated sentences evaluated against professional reference translations using modified n-gram precision

Comparing HT and MT Precision (2/2)



Decaying precision with increasing n

A point about length of n-grams

- 1 and 2-grams stress vocabulary match or lexical goodness

- 3-4 and higher n-grams stress structural match or syntactic goodness

Multiwords

Compositional

Meaning = Combination of meanings of parts

Also called as **collocations**

"Strong opposition"

Non-compositional

Meaning = Meaning cannot be made out from the meanings of parts

"White elephant"

‘Recall’ for MT Evaluation (1/2)

Case of Candidates shorter than references

English: *Will blue be able to understand quality of long sentence?*

Reference: क्या ब्लू लंबे वाक्य की गुणवत्ता को समझ पाएगा?

kya blue lambe vaakya ki guNvatta ko samajh paaega?

Candidate: **लंबे वाक्य**

lambe vaakya

long sentence

long sentence

modified unigram precision: 2/2 = 1

modified bigram precision: 1/1 = 1

Recall (2/2)

Reference 1: मैने खाना खाया

maine khaana khaaya

I food ate

I ate food

Candidate 2: मैने खाना खाया

maine khaana khaaya

I food ate

I ate food

P1= 1

Candidate longer than references

Reference 2: मैने भोजन किया

maine bhojan kiyaa

I meal did

I had meal

Candidate 1: मैने खाना भोजन किया

maine khaana bhojan kiya

I food meal did

I had food meal

P1: ¾=0.75

Formulating BLEU (Step 3): Incorporating recall

- Sentence length indicator of ‘good match’

- Brevity penalty (BP):
  - Multiplicative factor
  - Candidate translations that match reference translations in length must be ranked higher

Candidate 1: लंबे वाक्य

lambe bakya

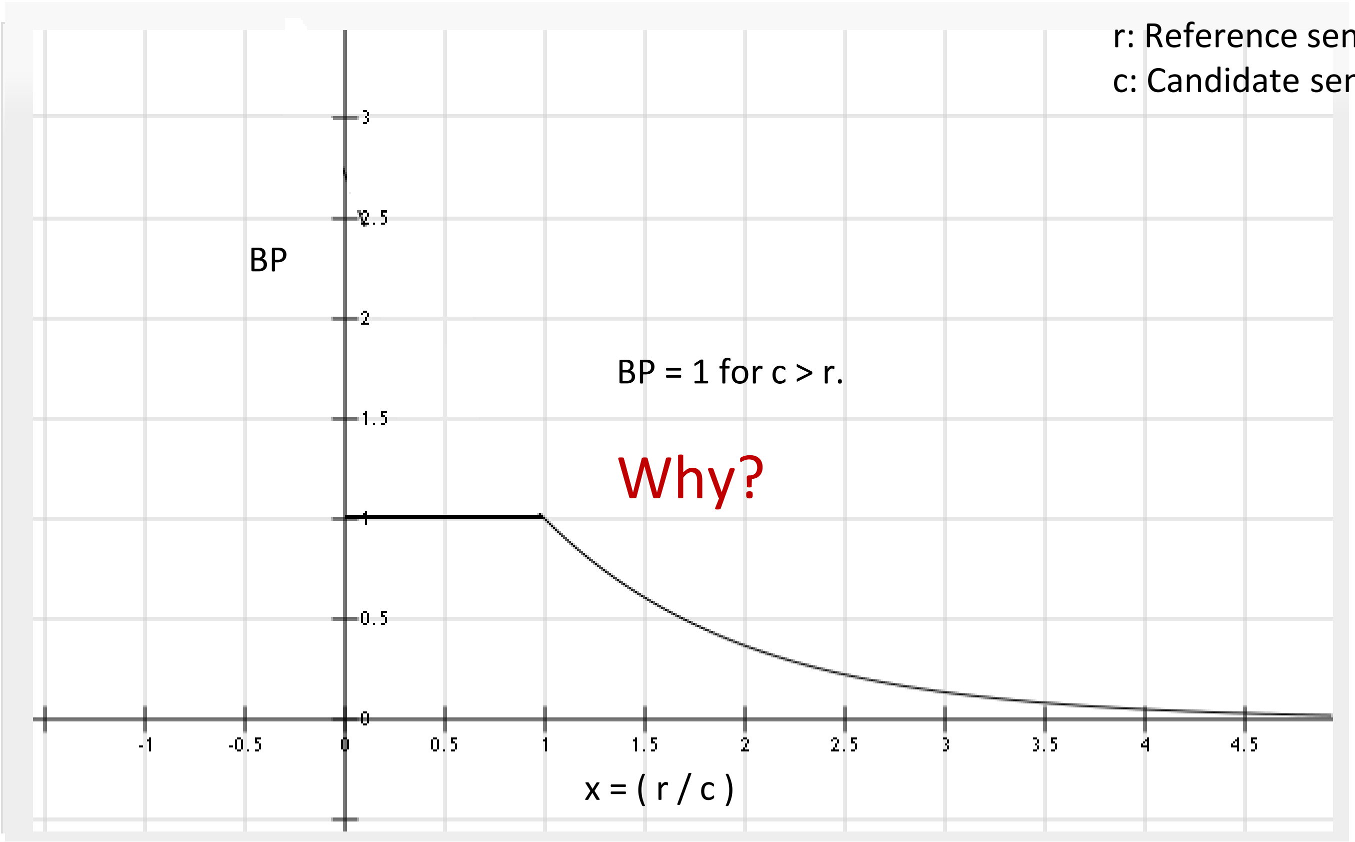
Candidate 2: क्या ब्लू लंबे वाक्य की गुणवत्ता समझ पाएगा ?

kya bleu lambe vakya ki gunvatta samajh payega ?

Formulating BLEU (Step 3): Brevity Penalty

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

r: Reference sentence length
c: Candidate sentence length



Graph drawn using www.fooplot.com

BP does not penalize translations longer than reference

Why?

Translations longer than reference are already penalized by modified precision

Validating the claim:

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

Final BLEU Score Formula

Recall -> Brevity Penalty

Precision→ Modified n-gram precision

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

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

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

Final BLEU Score Formula

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

N: The maximum n-gram length considered for precision matching (usually 4 or 5)

w_n: Weight for each n-gram precision, typically set to 1/N

p_n: Precision for each n-gram length.

Computing BLEU: Candidate-1

English: I had lunch now

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Reference:
मैने अभी खाना खाया
maine abhi khana khaya

BP= 1, for all n, w_n=1/2

Candidate 1:
मैने अब **खाना खाया**
maine ab khana khaya

BLEU=sqrt(0.75 X 0.33)=0.49

P1: 3/4 = 0.75,
P2: 1/3= 0.33

Computing BLEU: Candidate-2

English: I had lunch now

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Reference: मैने अभी खाना खाया
maine abhi khana khaya

BP= 1, for all n, w_n=1/2

Candidate 2: **मैने अभी** लंच एट
maine abhi lunch ate

BLEU=sqrt(0. 5 X 0.33)=0.40

Unigram precision: 2/4 = 0.5
Similarly, bigram precision: 0.33

Giving importance to Recall: Ref n-grams

ROUGE

- **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
- ROUGE is a package of metrics: ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S

ROUGE-N

$$\text{ROUGE-N} = \frac{\sum_{S\in\{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\boxed{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}}$$

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

ROUGE-N incorporates Recall

Will BLEU be able to understand quality of long sentences?

Reference translation:
क्या ब्लू लंबे वाक्य की गुणवत्ता को समझ पाएगा?
Kya bloo lambe waakya ki guNvatta ko samajh paaega?

Candidate translation:
लंबे वाक्य
Lambe vaakya

ROUGE-N: 1 / 8
Modified n-gram Precision: 1

Other ROUGEs

- ROUGE-L
  - Considers longest common subsequence
- ROUGE-W
  - Weighted ROUGE-L: All common subsequences are considered with weight based on length
- ROUGE-S
  - Precision/Recall by matching skip bigrams

ROUGE v/s BLEU

	ROUGE	BLEU
Handling incorrect words	Skip bigrams, ROUGE-N	N-gram mismatch
Handling incorrect word order	Longest common sub-sequence	N-gram mismatch
Handling recall	ROUGE-N incorporates missing words	Precision cannot detect ‘missing’ words. Hence, brevity penalty!

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Test of hypothesis

Terminology

A Practical Problem

- A bridge is being built. The weight it can tolerate has a distribution with $\mu=400$ and $\sigma=40$. A car that goes on the bridge has weight distribution given by $\mu=3$ and $\sigma=0.3$. We want the probability of damage to the bridge to be less than *0.1*. How many cars can we allow to go on the bridge?

When does the bridge break?

$$W_{total} > W_{tolerance}$$

Deterministic

- Damage if

$$3N=400$$

$$\Rightarrow N=133$$

Deterministic, but with bounds (1/2)

- Strongest bridge and lightest car
- Bridge withstand 440 and car weight 2.7
- Most liberal situation also most risky!

ceiling (2.7N=440)

$\Rightarrow$ N=163 !!

Deterministic, but with bounds (2/2)

- Weakest bridge and heaviest car
- Bridge withstand 360 and car weight 3.3
- Most conservative situation and safest
- But resource wise most inefficient!!

floor(3.3N=360)

$\Rightarrow$ N=109 !!

Lets look at these numbers for a while

- Most liberal, 163 nos.
- Most conservative, 109 nos.
- What should be the ACTUAL NO. of cars to be allowed?
- This is an OBJECTIVE DECISION
- A precise no. has to be allowed
- How much is that?

Depends on the priority: safety the only consideration

- As an Administrator, I want to PLAY VERY SAFE
- No risk
- Then only 109 cars
- Bridge will never break
- I am safe

Point of view and priority: earning first, throughput first, efficiency first

- I want to have maximum utilization of the bridge
- Maximum earning from toll
- Maximum movement across river
- Maximum economic activity
- Maximum interaction
- People happy ☺

But risk is higher!

- The bridge will VERY LIKELY cross the tolerance limit
- Bridge breaks
- Lives lost
- Property damaged
- People unhappy ☹

Relate to covid-19 situation?

- Yes
- Do not go out
- Do not interact
- Very safe
- But no economic and social activity
- How to sustain?
- How to break monotony

Need balance, sweet spot is somewhere in between, MIDDLE PATH

How to get the sweet spot? The middle path?

- Answer

PROBABILITY

Back to the bridge

- MOO: Multi-objective Optimization
- Many objectives to be satisfied
  - Safety
  - Utilization of facility
  - Earning
  - People satisfaction
  - *Etc.*

Bring in probability

- #cars = N
- Each car's weight is normal with $\mu=3$ and $\sigma=0.3$
- Invoke Central Limit Theorem