CS772: Deep Learning for Natural Language Processing

Evaluation

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week 15 of 11th April, 2022

NLP evaluation

Focus on MT evaluation

(Credit: Aditya Joshi, Kashyap Popat, Shubham Gautam)

Precision/Recall

Precision:

How many results returned were correct?

Recall:

What portion of correct results were returned?

Adapting precision/recall to NLP tasks

Document Classification: Taxonomy

- Labels form a taxonomy
- E.g.
  - Financial
    - Stocks
    - Tradings
    - Merger and acquisition, etc.
  - Sports
  - Cultural
  - Literature

Document Retrieval and Classification

- **Document Retrieval**

Precision =

$$\frac{|\text{Documents relevant and retrieved}|}{|\text{Documents retrieved}|}$$

Recall=

$$\frac{|\text{Documents relevant and retrieved}|}{|\text{Documents relevant}|}$$

- **Classification**

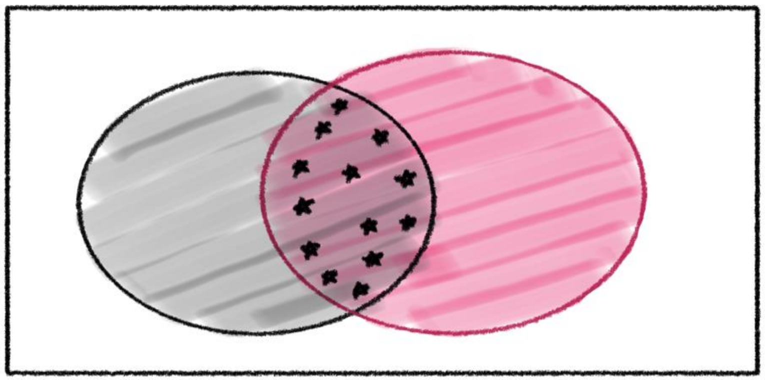
Precision =

$$\frac{|\text{True Positives}|}{|\text{True Positives + False Positives}|}$$

Recall=

$$\frac{|\text{True Positives}|}{|\text{True Positives + False Negatives}|}$$

Venn Diagram illustrating "Actual" vs "Obtained"



SET S_1 --- OBTAINED

SET S_2 --- ACTUAL

$S_1 \cap S_2$ ... TRUE POSITIVES

$S_1 - (S_1 \cap S_2)$ ... FALSE POSITIVES

$S_2 - (S_1 \cap S_2)$ ... FALSE NEGATIVES

$(S_1 \cup S_2)^c$ ... TRUE NEGATIVES

$$Precision = \frac{|S_1 \cap S_2|}{|S_1|}$$

$$Recall = \frac{|S_1 \cap S_2|}{|S_2|}$$

Type 1 and Type 2 errors

	False Positive	False Negative
statistical hypothesis testing	A **type I error** is the rejection of a of a true null hypothesis e.g. "an innocent person is convicted"	A **type II error** is the non-rejection of a false null hypothesis e.g. "a guilty person is not convicted"
Philosophy logic and language	Error of commission	Error of omission

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

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

Reading

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.

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

Candidate 2: **मैने अभी** लंच एट

maine abhi lunch ate.

I now lunch ate

I ate lunch(OOV) now(OOV)

matching unigrams: 2,

matching bigrams: 1

Unigram precision: Candidate 1: 3/4 = 0.75, Candidate 2: 2/4 = 0.5

Similarly, bigram precision: Candidate 1: 0.33, 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 (Step 2): Modified Precision

- Clip the total count of each candidate word with its maximum reference count
- Countclip(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

Modified precision for n-grams

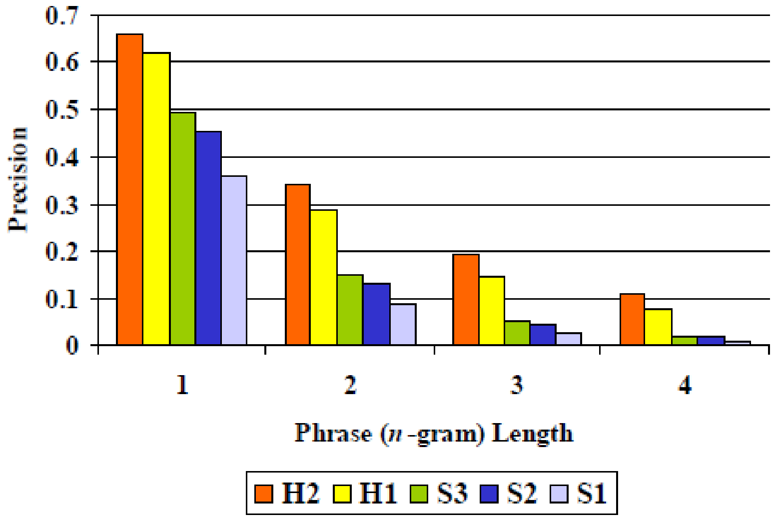
Overall candidates of test corpus

n-gram': All n-grams in C

Calculating modified n-gram precision (1/2)

- 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

Calculating modified n-gram precision (2/2)



- Decaying precision with increasing n
- Comparative ranking of the five

Combining precision for different values of n-grams?

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

Formulation of BLEU: Recap

- Precision cannot be used as is
- Modified precision considers ‘clipped word count’

‘Recall’ for MT (1/2)

- Candidates shorter than references
- Reference: क्या ब्लू लंबे वाक्य की गुणवत्ता को समझ पाएगा?

  kya blue lambe vaakya ki guNvatta ko samajh paaega?

  will blue long sentence-of quality (case-marker) understand able(III-person-male-singular)?

  Will blue be able to understand quality of long sentence?

Candidate: **लंबे वाक्य**

lambe vaakya

long sentence

long sentence

modified unigram precision: 2/2 = 1

modified bigram precision: 1/1 = 1

Recall for MT (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

Modified unigram precision: 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

Modified unigram precision: 1

Formulating BLEU (Step 3): Incorporating recall

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

Candidate 1: लंबे वाक्य

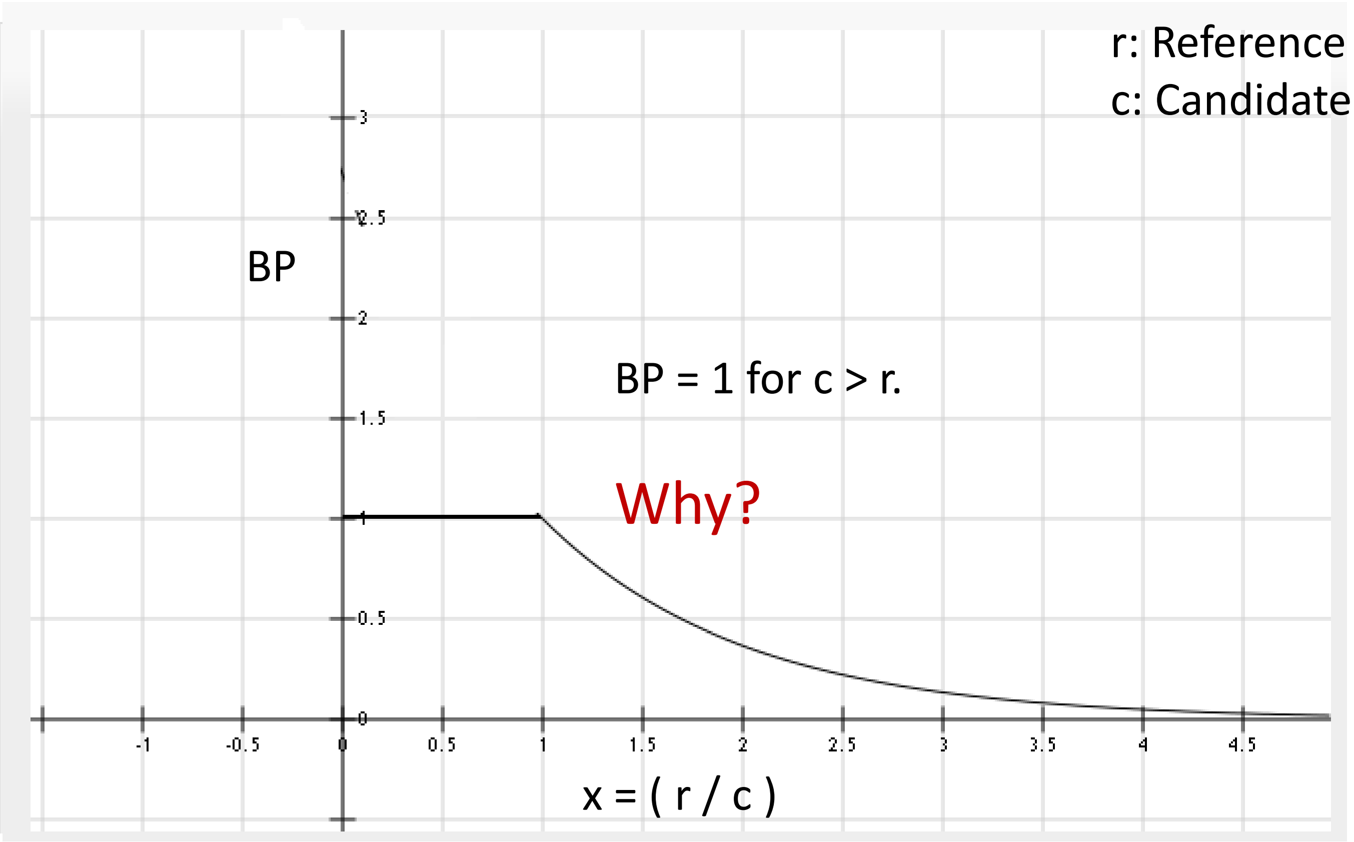
Candidate 2: क्या ब्लू लंबे वाक्य की गुणवत्ता समझ पाएगा?

Formulating BLEU (Step 3): Brevity Penalty

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

r: Reference sentence length

c: Candidate sentence length



Graph drawn using www.fooplot.com

BP leaves out longer translations

Why?

Translations longer than reference are already penalized by modified precision

Validating the claim:

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

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

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

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

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

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

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

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 (suite of metrics)	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!

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

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

Course summary (1/3)

- Week 1: Introduction
  - Ambiguity
  - Data, ML and Disambiguation

- Week 2, 3: Language Word Modelling, Word Vectors, Skip Gram
  - LM
  - Skip gram
  - Perceptron

- Week 4,5, 6
  - FFNN, BP
  - Word vector n/w training
  - Softmax and Cross Entropy

Course summary (2/3)

- Week 7, 8, 9: RNN
  - BPTT
  - Hopfield Net and Boltzmann Machine
  - LM through RNN

- Week 10, 11: CNN
  - Kernels or filters
  - Applications of CNN in Vision
  - Applications in NLP

Course summary (3/3)

- Week 12, 13, 14: Attention, Transformer, and Transformer Applications
  - Importance of Attention in NLP: subject-verb agreement, wsd, coreference
  - Stackes of Encoder-Decoder layers in Transformers
  - Application in MT
  - Application in NLG

- Week 15: Evaluation in NLP
  - Precision and Recall
  - BLEU
  - ROUGE