Automatic Machine Translation Evaluation

Koel Dutta Chowdhury
UdS

Summer Semester 2023
18 April 2023

MT Evaluation

Recap: Manual evaluation

- Scoring, ranking, diagnostic,..,intrinsic, extrinsic etc.
- Time consuming
- Expensive
- Difficult to define and operationalise
- Hard to reproduce: low inter- and intra-rater agreement
- Hard to scale: crowdsourcing

Recap:Automatic evaluation

F-Measure

Reference:

`Israeli officials are responsible for airport security.`

System A:

`Israeli officials responsibility of airport safety.`

System B:

`airport security Israeli officials are responsible.`

System C:

`security Israeli are officials responsible airport.`

- **Precision**: how many of the words in output are correct (in ref)?
- **Recall**: how many of the words in reference are in the output?
- **F-score**: harmonic mean of precision and recall

Automatic evaluation

F-Measure

	System A	System B	System C
precision	0.50	1.00	1.00
recall	0.43	0.86	0.86
f-score	0.46	0.92	0.92

- f-measure can reward unintelligible word salad in C if individual words are OK
- Fails to reflect word order!

Automatic Evaluation: BiLingual Evaluation Understudy

BLEU, Papineni, Roukos, Ward and Zhu (2001)

Ref: `Israeli officials are responsible for airport security.`

Sys A: `Israeli`

Sys C: `Israeli officials responsibility of airport safety.`

Look at n-gram overlap, not just word overlap

n-gram precision $n = 1...4$ times a brevity penalty ("recall")

BLEU: =

$$\min\left(1, \exp\left(1 - \frac{|\text{reference}|}{|\text{output}|}\right)\right) \cdot \left(\prod_{n=1}^{4} n - gram \quad precision\right)^{\frac{1}{4}}$$

BLEU$=0$ if the hypothesis does not have at least one matching n-gram for any one of the n-gram precision $n = 1...4$: systems A and C!

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Reference:

`Israeli officials are responsible for airport security.`

System B:

`airport security Israeli officials are responsible.`

- BLEU: =

$$\min\left(1, \exp\left(1 - \frac{|\text{reference}|}{|\text{output}|}\right)\right) \cdot \left(\prod_{n=1}^{4} n - gram \quad precision\right)^{\frac{1}{4}}$$

$$\left(\prod_{n=1}^{4} n - gram \quad precision\right)^{\frac{1}{4}} = \left(\frac{6}{6} \times \frac{4}{5} \times \frac{2}{4} \times \frac{1}{3}\right)^{\frac{1}{4}} = 0.60$$

$$\min\left(1, \exp\left(1 - \frac{|\text{reference}|}{|\text{output}|}\right)\right) = \min\left(1, \exp\left(1 - \frac{7}{6}\right)\right) = 0.87$$

$$BLEU = 0.87 \times .60 = 0.52$$

Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

- Problem: BLEU assigns 0 to many hypotheses ...
- Meant to work on document, not individual sentence level
- sBLEU for sentence level ... (smoothed BLEU)

Fancy way of writing BLEU

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

- BP: Brevity penalty
- Taking log of n-gram precision (P_n), summing over them and using inverse function of log
- w_n positive weights summing to one

Automatic evaluation

IBM BLEU vs. NIST BLEU vs. ...

There are several widely used implementations of BLEU.

(Moses `multi-bleu.perl` script, NIST `mteval-vXX.pl` script, etc.)

Results differ because of:

- Different tokenisation schemes.
- Different definition of *closest reference* in the brevity penalty estimation.
  $\rightarrow$ SacreBLEU [Pos18]

Criticisms of BLEU

- Brevity penalty is not a good measure of recall
- Do not consider global grammaticality
- Punishes perfect paraphrases: Do not consider meaning
  - Yesterday John resigned from the company
  - John quit the company yesterday
- Geometric ngram averaging is volatile to "zero" scores
- Requires exact word matches, but not stemmed word matches, synonym and semantically-related word matches

Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P***recision** and *R***ecall** weighted harmonic mean

$$Pen = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}}\right)^\beta$$

Penalty factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

$$F_\alpha = \frac{PR}{\alpha P + (1-\alpha)R}$$

*P***recision** and *R***ecall** weighted harmonic mean

$$Pen = \gamma \left(\frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta$$

Penalty factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase

METEOR: Flexible Matching

- Explicitly aligns the words in the MT output with their corresponding matches in the reference translations
- Exact module: maps two words if they are exactly the same.
- Porter stem module: maps two words if they are the same after they are stemmed using the Porter stemmer
  - Partial credit for matching stems
    - ★ system : Jim walk home
    - ★ reference : Joe walks home
- WN synonymy module: maps two words if they are considered synonyms, based on the fact that they both belong to the same synset in WordNet.
  - Partial credit for matching synonyms
    - ★ system:Jim strolls home
    - ★ reference:Joe walks home

Automatic Evaluation
TER, [SDS+06]

Translation Edit Rate, TER

$$TER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}$$

REF: Saudi Arabia denied this week information published in the American New York Times

HYP: This week the Saudis denied information published in the New York Times

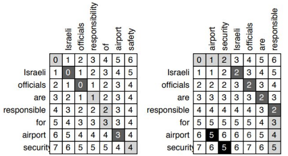
- Insertion: American
- Shifts: this week
- Substitutions: SAUDI ARABIA vs. THE SAUDIS

$$TER = \frac{1 + 2 + 1}{13} = 31\%$$

Automatic Evaluation

Other error rates

- WER : Word Error Rate
  - Minimum number of editing steps to transform output to reference



Metric	System A	System B
word error rate (WER)	57%	71%

PER : Position-independent Word error Rate

  - Does not consider the order of words

SER : Sentence Error Rate

  - The percentage of sentences that are identical to reference sentences

chrF, Popovic (2015)

Character F-score

$$\text{CharF}_\beta = \frac{(1+\beta^2)\text{CharP} \cdot \text{CharR}}{\beta^2 \cdot \text{CharP} + \text{CharR}}$$

- CharP is the character precision, which is the proportion of characters in the output text that also appear in the reference text.
- CharR is the character recall, which is the proportion of characters in the reference text that also appear in the output text.
- β is a parameter that controls the trade-off between precision and recall.
- Typically, β is set to 1 to give equal weight to precision and recall.
- if β is more than 1, recall component is being weighted more relative to the precision component

chrF vs BLEU

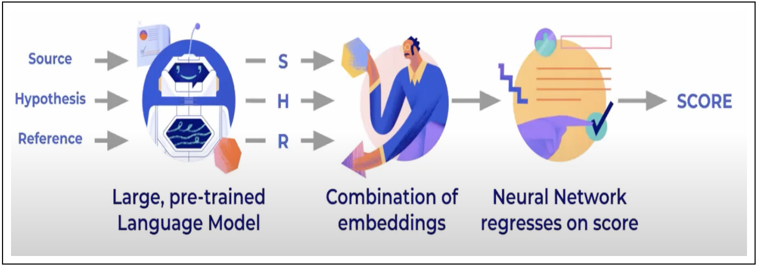
- Measures character n-gram overlap instead of word n-grams as in BLEU.
- Reduces sensitivity to sentence tokenisation
- Useful for tasks where word boundaries may not be well-defined
- Character sequences matching helps in recognizing different forms of a single word.
- Assigns partial reward for incorrectly spelled words.

COMET, Rei (2020)

Cross-lingual Optimized Metric for Evaluation of Translation

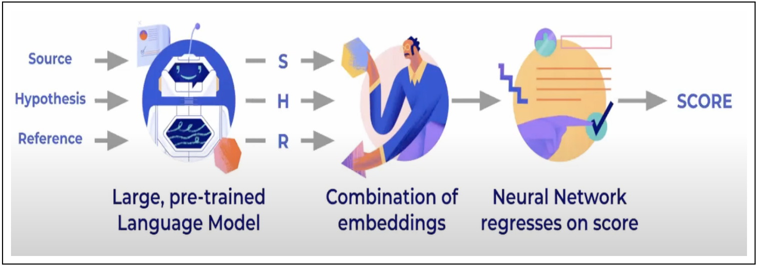
Important: exact lexical matching is a crude estimate for sentence level similarity in meaning.
$\rightarrow$ fails to recognise and capture semantic similarity!

COMET: Basic Modeling Approach



- Levarages a pre-trained multilingual model and is trained to mimic human ratings of translations
- Consists of an Estimator and Translation Ranking Model based on human determination
- Estimator is trained to regress directly on a quality score, the Translation Ranking model is trained to minimize the distance between a better hypothesis and its reference (or input original source)
- Correlates well with different types of human judgements

COMET: Basic Modeling Approach



- Levarages a pre-trained multilingual model and is trained to mimic human ratings of translations
- Consists of an Estimator and Translation Ranking Model based on human determination
- Estimator is trained to regress directly on a quality score, the Translation Ranking model is trained to minimize the distance between a better hypothesis and its reference (or input original source)
- Correlates well with different types of human judgements

Evaluation and Key Performance Numbers

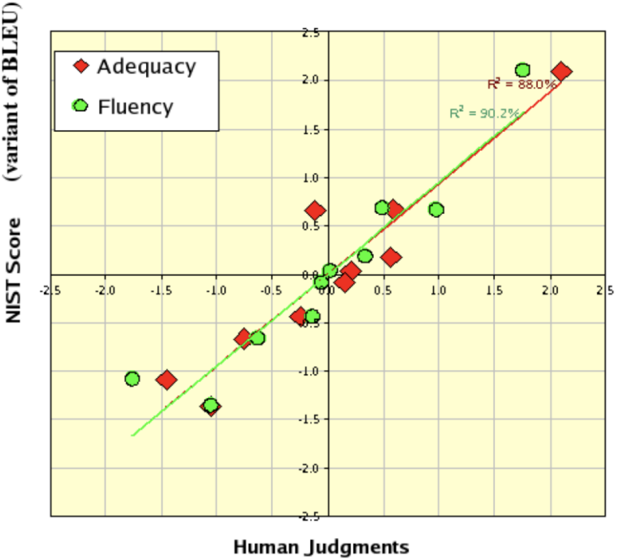
	Language		Reference	Method		Rank
	any	pretrained	needed	string	embeddings	@WMT22
BLEU	✓	✗	✓	n-gram match	✗	19
chrF	✓	✗	✓	n-gram match	✗	16
TER	✓	✗	✓	edit distance	✗	–
COMET	✗	✓	✓	✗	src,hyp,ref	2, 5
UniTE	✗	✓	✓	✗	src,hyp,ref	3
BleuRT	✗	✓	✓	✗	hyp,ref	4
BertScore	✗	✓	✓	✗	hyp,ref	14
COMETKiwi	✗	✓	✗	✗	src,hyp	7
UniTE-Src	✗	✓	✗	✗	src,hyp	9

Table 5.2: Representative MT automatic evaluation metrics and their ranking in the last WMT Metrics Shared Task.

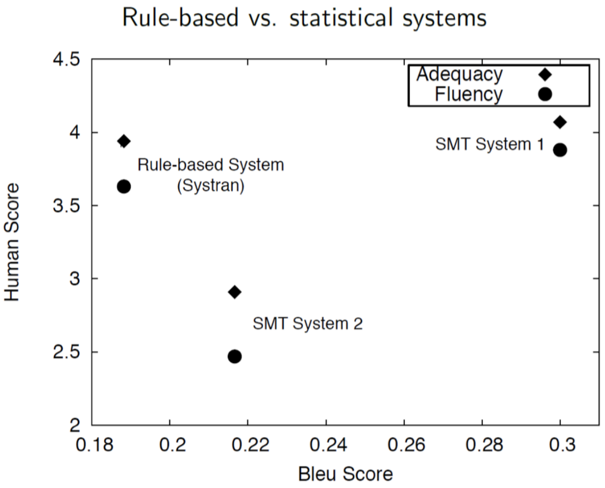
Evaluation of Evaluation Metrics

- Automatic metrics are low cost, tunable, consistent
- But are they correct?
  $\rightarrow$ Yes, if they correlate with human judgement

Correlation with Human Judgement



Correlation with Human Judgement



Correlations of metrics with human ranking

Metric	de-en	en-de
BLEU	.88	.76
WER	.93	.82
PER	.84	.73
ChrF1	.93	.87
ChrF3	.96	.90
Beer	.95	.88

(System level, WMT 2017)

- Evaluation is important in the system development cycle. Automatic evaluation accelerates significatively the process.
- Human evaluation is expensive. Automatic evaluation is cheap, but not always fair
- Active development of new metrics goes beyond lexical similarity.
  - syntactic similarity
  - semantic equivalence or entailment
  - metrics targeted at reordering
  - trainable metrics
- Evaluation campaigns that rank metrics
- Be careful when you argue about MT quality!

Baseline metrics and participants of WMT21 Metrics Shared Task

	Metrics	broad category	Citation	Availability
Baselines	sentBLEU	lexical overlap	Papineni et al. (2002)	https://github.com/mjpost/sacrebleu
	BLEU	lexical overlap	Papineni et al. (2002)	https://github.com/mjpost/sacrebleu
	TER	lexical overlap	Snover et al. (2006)	https://github.com/mjpost/sacrebleu
	chrF	lexical overlap	Popović (2015)	https://github.com/mjpost/sacrebleu
	BERTscore	embedding similarity	Zhang et al. (2020)	https://github.com/Tiiiger/bert_score
	PRISM	MT-model-based	Thompson and Post (2020)	https://github.com/thompsonb/prism
Participants	COMET-*	neural finetuned metrics	Rei et al. (2021)	https://github.com/Unbabel/COMET
	OpenKiwi-MQM	neural finetuned metrics	Kepler et al. (2019)	https://github.com/Unbabel/OpenKiwi
	YiSi-*	embedding similarity	Lo (2019)	https://github.com/nrc-cnrc/yisi
	MTEQA	question-answer	Krubiński et al. (2021a)	https://github.com/ufal/MTEQA
	RegEMT-*	Ensemble	Stefanik et al. (2021)	https://github.com/MIR-MU/regemt
	RoBLEURT	neural finetuned metrics	Wan et al. (2021)	Not a public metric
	BLEURT-*	neural finetuned metrics	Sellam et al. (2020)	https://github.com/google-research/bleurt
	cushLEPOR-*	lexical overlap	Han et al. (2021)	https://github.com/poethan/cushLEPOR
	C-SPEC-*	neural finetuned metrics	Takahashi et al. (2021)	Not a public metric
	MEE-*	lexical and embedding similarity	Mukherjee et al. (2020)	https://github.com/AnanyaCoder/MEE_WMT2021

References

- BLEU [PRWZ02]
- NIST [Dod02]
- METEOR [BL05]
- ROUGE [LO04]
- CharF [Pop15]
- COMET [RSFL20]

References

- GTM [MGT03]
- BLANC [Dod02]
- BertScore [ZKW+19]
- BLEURT [**?**]
- CDER [LUN06]
- ULC [GA06]

References I

Satanjeev Banerjee and Alon Lavie.
METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.
In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

George Doddington.
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145, 2002.

Jesús Giménez and Enrique Amigó.
IQMT: A Framework for Automatic Machine Translation Evaluation.
In *Proceedings of the 5th LREC*, pages 685–690, 2006.

Chin-Yew Lin and Franz Josef Och.
Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics.
In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

References II

Gregor Leusch, Nicola Ueffing, and Hermann Ney.
CDER: Efficient MT Evaluation Using Block Movements.
In *Proceedings of EACL*, pages 241–248, 2006.

I. Dan Melamed, Ryan Green, and Joseph P. Turian.
Precision and Recall of Machine Translation.
In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.

Maja Popović.
chrf: character n-gram f-score for automatic mt evaluation.
In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

Matt Post.
A call for clarity in reporting bleu scores.
arXiv preprint arXiv:1804.08771, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the Association of Computational Linguistics*, pages 311–318, 2002.

References III

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie.
Comet: A neural framework for mt evaluation.
arXiv preprint arXiv:2009.09025, 2020.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul.
A study of translation edit rate with targeted human annotation.
In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi.
Bertscore: Evaluating text generation with bert.
arXiv preprint arXiv:1904.09675, 2019.