Think Inside the Box: Glass-box Evaluation Methods for Neural MT

Marina Fomicheva

INLG
18 December 2020

Machine Translation as an NLG task

- NLG converts a meaning representation into a NL utterance
- The focus of this talk is on MT evaluation
- How is it different?
  - MT is constrained by the original sentence
  - But still a lot of potential variability in the space of possible outputs
    - Underspecification and ambiguity
    - Lack of extra-sentential and extra-linguistic context

What makes MT evaluation challenging?

- Large space of possible correct translations
- Multiple different aspects involved in evaluation
  - Definition of quality
    - Adequacy/fluency scales, preference judgements
    - Error annotation
    - Task-oriented evaluation (e.g. PE effort)
  - Granularity
    - System-level vs. sentence-level
    - Document level -> sentence level -> word level

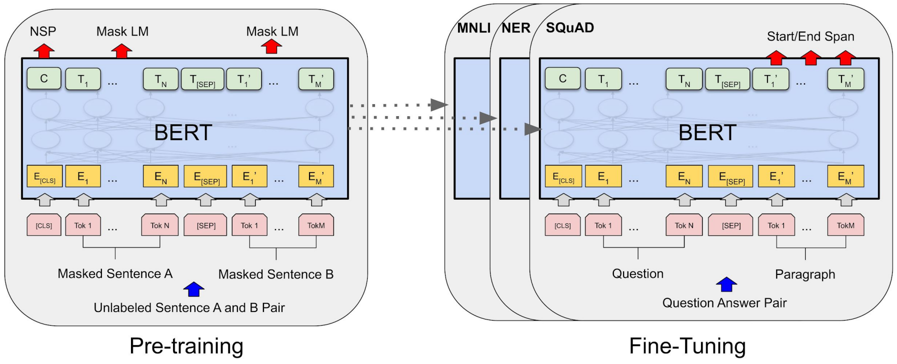
Approaches to Automatic MT Evaluation

	Automatic evaluation	**Quality estimation**
Input representation	MT output Human reference(s)	Source MT output
Learning mechanism	No	Yes
Supervision	Human reference(s)	Gold quality labels
Algorithm	Similarity metrics	Feature-based ML
MT system	Black-box	Black-box/Glass-box (statistical MT)
Gold labels	Intrinsic quality measures	Task-oriented, e.g. HTER
Meta-evaluation metric	Spearman/Pearson correlation	RMSE/MAE

Approaches to Automatic MT Evaluation

	Automatic evaluation	Quality estimation
Input representation	Source, MT output, Reference(s)	Source, MT output **MT system**
Learning mechanism	Yes	
Supervision	Reference(s) Gold quality labels Pseudo-references **MT hypotheses**	Gold quality labels **Unsupervised**
Algorithm	Similarity metrics	Feature-based ML
	NN-based systems Pre-trained representations (BERT)	
MT system	Black-box/**Glass-box (neural MT)**	
Gold labels	Intrinsic quality measures/HTER	
Meta-evaluation metric	Spearman/Pearson correlation	

Approaches to Automatic MT Evaluation

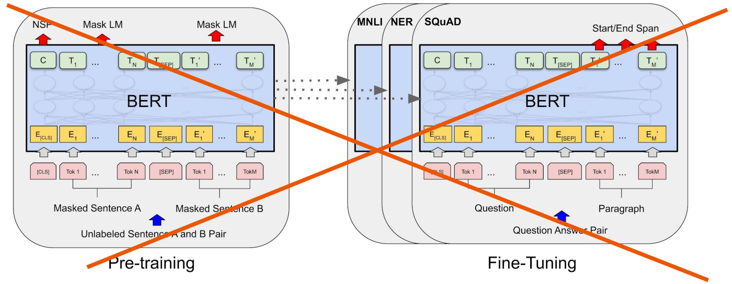


Pre-trained contextualised multilingual representations [Devlin et al. 2018, Conneau et al. 2019]

- Most recent SOTA in MT evaluation and QE
  - BertSCORE [Zhang et al., 2019]
  - Winning submissions to WMT2020 QE Shared Task [Fomicheva et al. 2020, Ranasinghe et al. 2020]
- Up to Pearson correlation of 0.9 with human judgments
- **But very resource-heavy models**

This work

- Bergamot project: https://browser.mt
  - Client-side MT in a web-browser
  - Alongside MT outputs, provide quality estimates
- Requirements for quality estimation
  - **Efficient: light and fast models**
  - Robust: open domain and language independent
  - Little or no supervision



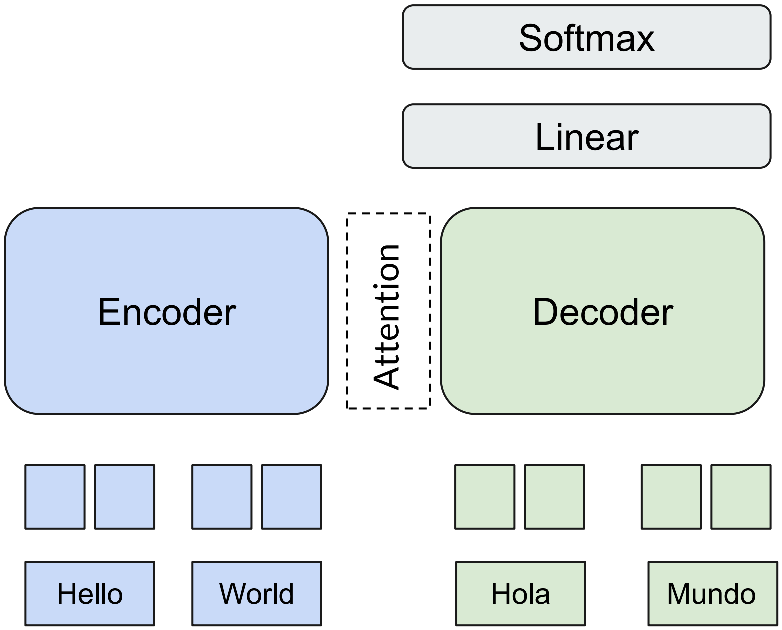
This talk: glass-box evaluation for NMT

- **What if instead of training neural models to evaluate MT quality we use the one we already have?**
- How to exploit internal information from the MT system
    - For quality estimation
    - For reference-based MT evaluation
- Assumption
    - If the model is confident then translation is good
    - How to measure confidence?

Glass-box Evaluation Methods for Neural MT

Fomicheva et al. (TACL2020). Unsupervised Quality Estimation for Neural Machine Translation
Fomicheva et al. (ACL2020). Multi-hypothesis Machine Translation Evaluation

NMT Reminder



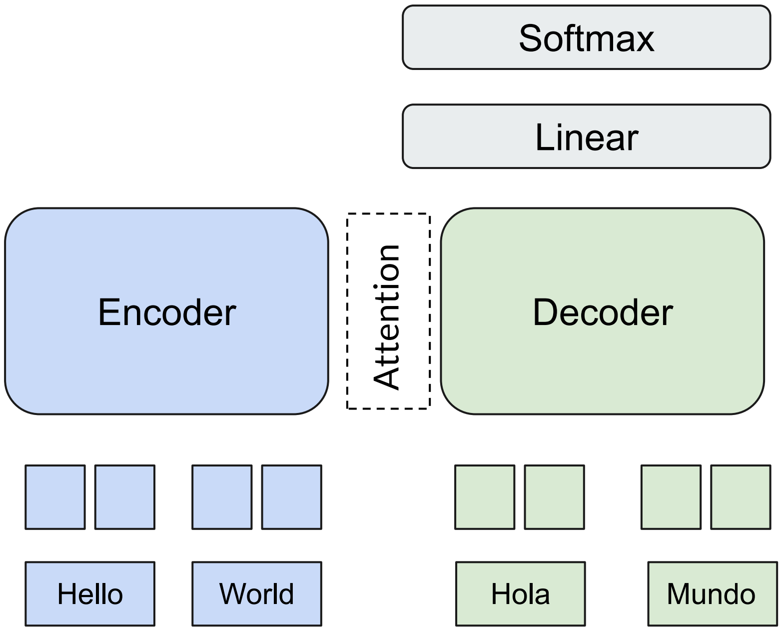
Assume seq-seq model with attention

Encoder maps the input sequence $x=x_1..x_N$ into a sequence of hidden states

Summarized into a single representation via attention mechanism

Given this representation, the decoder produces an output sequence $y=y_1..y_T$ one word at a time

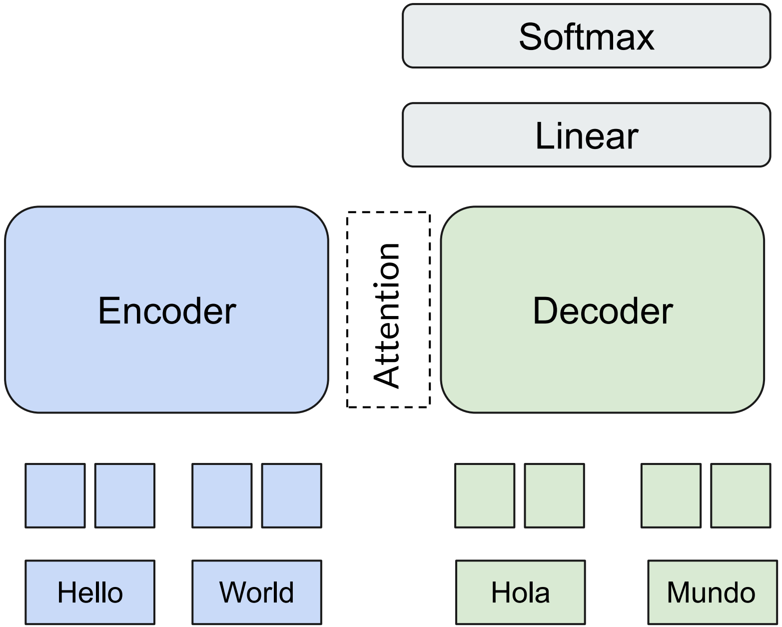
NMT Reminder



Linear layer projects decoder output into a logits vector $\in \mathbb{R}^{\mathcal{V}}$ where $\mathcal{V}$ is the size of target vocabulary

Softmax layer turns logits into probabilities

NMT Reminder

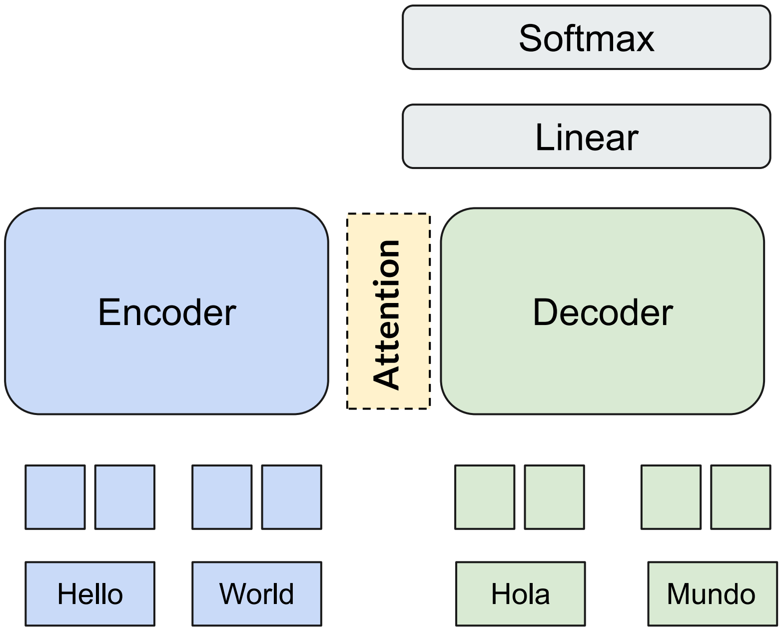


At each time step the decoder produces a conditional probability distribution over all the words in $\mathcal{V}$

$$p(\mathbf{y}|\mathbf{x},\theta) = \prod_{t=1}^{T} p(y_t|\mathbf{y}_{<t},\mathbf{x},\theta)$$

The word with the highest probability is returned as output at given time step

Glass-box evaluation for NMT



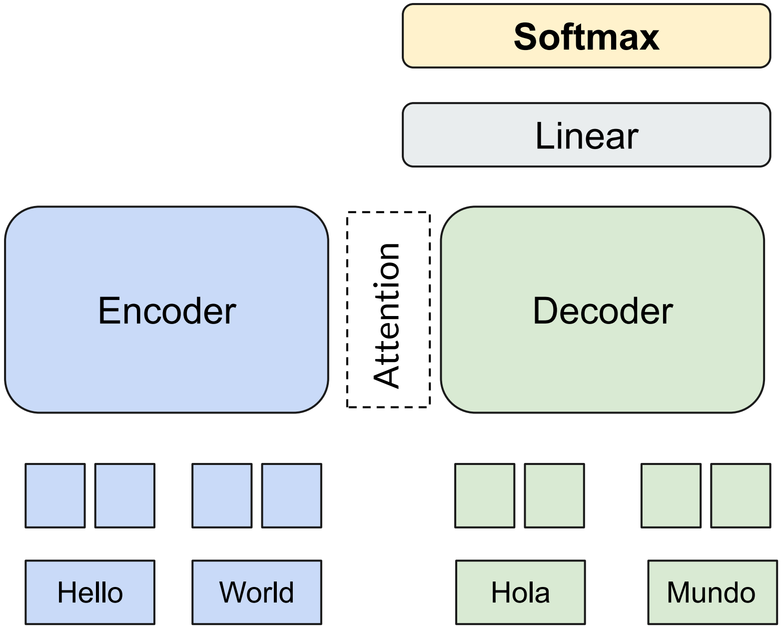
Encoder-decoder attention

Strength of connection between source and target tokens as an indicator of confidence

Entropy of encoder-decoder attention weights

$$\text{Att-Ent} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_{ji}\log\alpha_{ji}$$

Glass-box evaluation for NMT



Output probability distribution

- Log-probability of predicted tokens
- Entropy of the softmax distribution
- Dispersion of token-level probabilities

Glass-box evaluation for NMT

- Log-probability of the predicted tokens
- Averaged to get a sentence-level estimate

This	is	a	phrase		
0.5	0.9	0.8	0.6		0.7

$$\mathrm{TP} = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t | \mathbf{y}_{<t}, \mathbf{x}, \theta)$$

This talk: glass-box evaluation for NMT

Entropy of the output distribution

This	is	a	phrase

$\mathcal{V}$

phrase	0.6
sentence	0.3
...	...
zzz	0.0001

$$\text{Softmax-Ent} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{v=1}^{V} p(y_t^v)\log p(y_t^v)$$

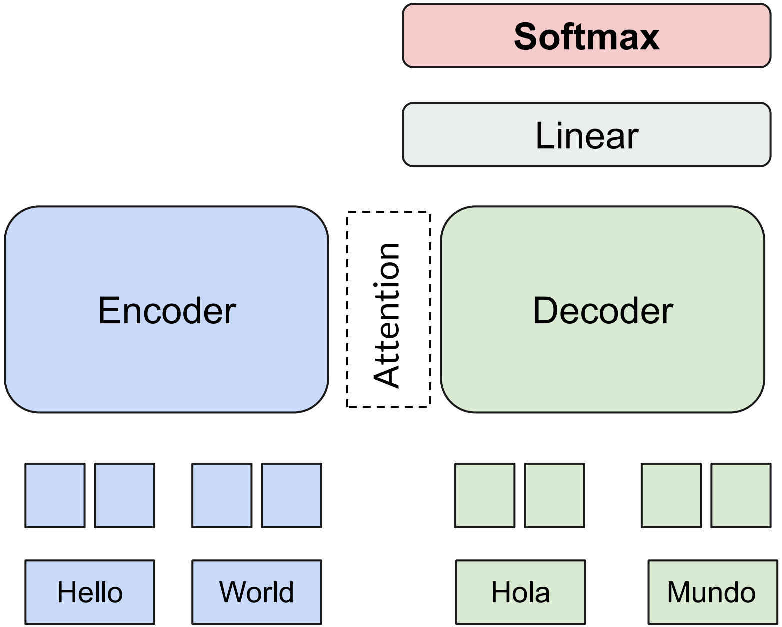
Glass-box evaluation for NMT

Dispersion of token-level probabilities

This	is	a	phrase
0.5	0.4	0.5	0.6
0.7	0.9	0.2	0.2

$$\text{Sent-Std} = \sqrt{\mathbb{E}[\mathrm{P}^2] - (\mathbb{E}[\mathrm{P}])^2}$$

Glass-box evaluation for NMT



Overconfident predictions

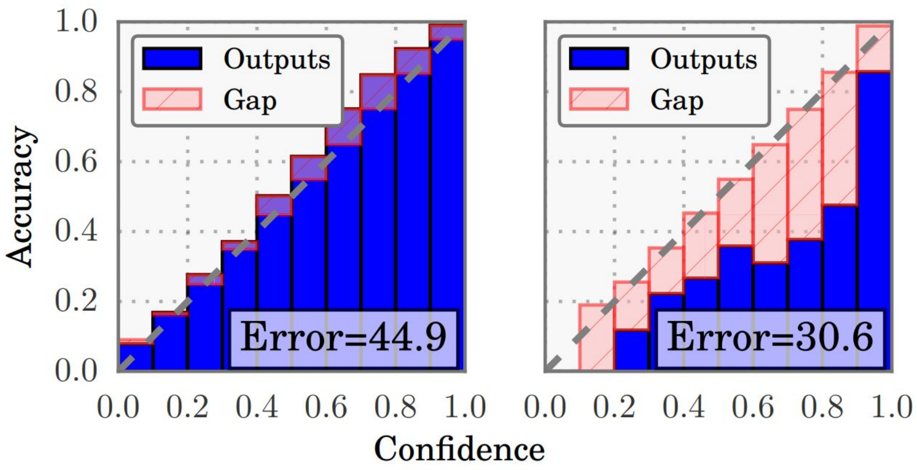
Neural networks can return wrong predictions with high probability

Softmax does not properly capture predictive uncertainty

- Aleatoric uncertainty (data)
- Model uncertainty (parameters)
- ...

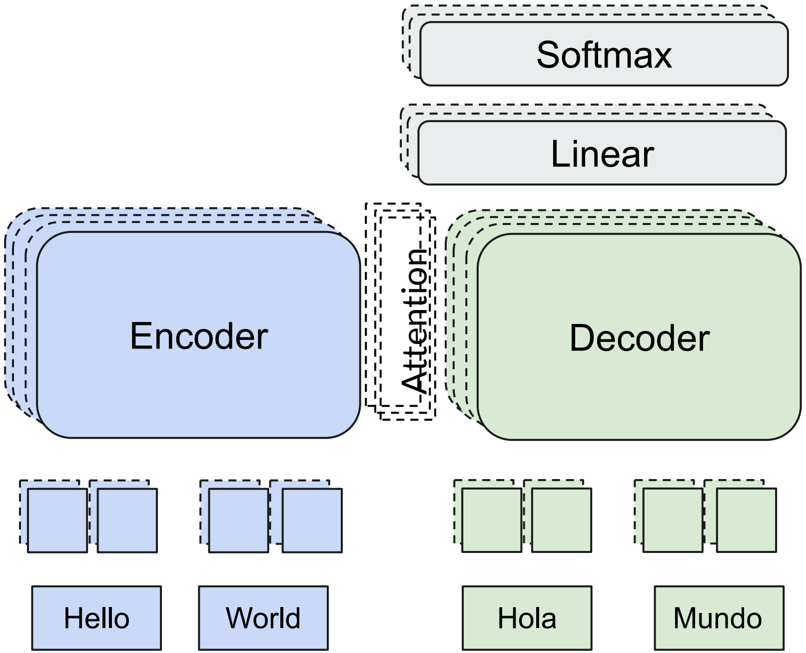
Glass-box evaluation for NMT

The problem of MT evaluation becomes the problem of calibration and uncertainty estimation in neural networks



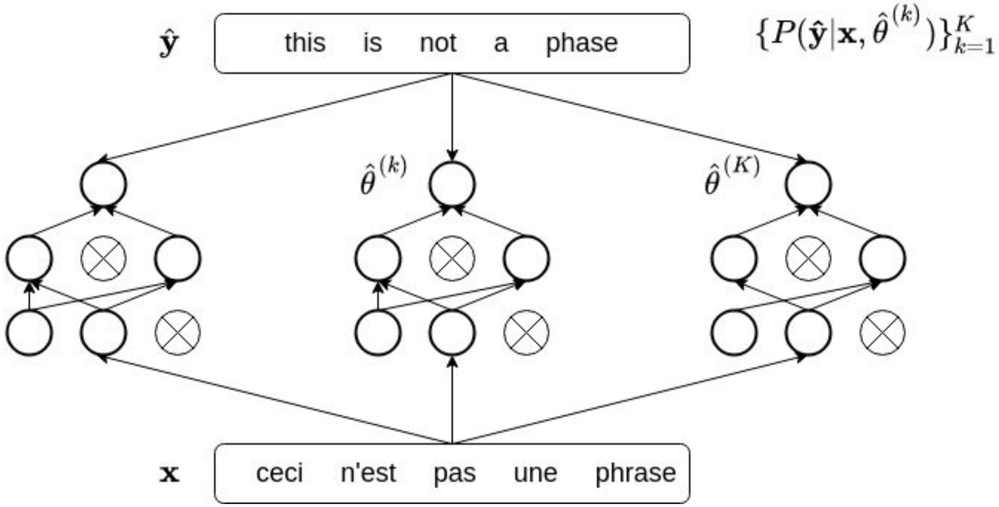
Guo et al. 2018

Glass-box evaluation for NMT



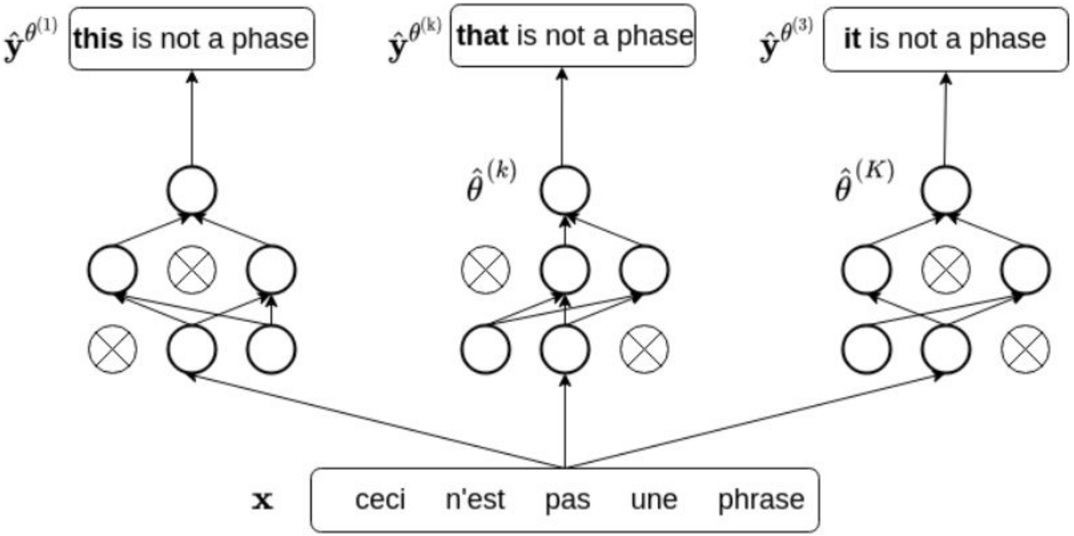
- Bayesian approach
  - Many possible models can explain the training data
  - Replace point estimates of model weights with probability distributions
- Prohibitive costs for deep NN
- Simpler approximations
  - **Monte Carlo Dropout** [Gal and Ghahramani, 2016]

MC Dropout for Quality Estimation



- Keep source and translation the same
- Compute segment-level translation probabilities K times with perturbed parameters
- Report mean and variance of the resulting distribution

MC Dropout for Quality Estimation



- Run inference K times with perturbed parameters
- Measure lexical similarity between generated translations

Example: High-Quality Estonian-English MT

Source	Siis aga võib tekkida seesmise ja välise vaate vahele l ohe.
Reference	This **could however** lead to a **split** between **inner and outer view**.
MT output 1-best	Then there **may be** a **split** between **internal and external viewpoints**.
MT hypotheses **MC Dropout**	Then, **however**, there **may be** a **split** between **internal and external viewpoints**.
	Then, however, there **may be** a **gap** between **internal and external viewpoints**.
	Then there **may be** a **gap** between **internal side and the external view**.
	Then there **may be** a **split** between **internal and external perspectives**.

Example: Low-Quality Estonian-English MT

Source	Tanganjikast püütakse niiluse ahvenat ja kapentat.
Reference	Nile perch and kapenta are fished from Lake Tanganyika.
MT output 1-best	There is a **silver thread and candle from Tanzeri**.
MT hypotheses MC Dropout	There will be a **silver thread and a penny from Tanzer**.
	There is an **attempt at a silver greed and a carpenter from Tanzeri**.
	There will be a **silver bullet and a candle from Tanzer**.
	The **puzzle** is being caught in the **chicken's gavel and the coffin**.

Example: Low-Quality Estonian-English MT

Source	Tanganjikast püütakse niiluse ahvenat ja kapentat.
Reference	Nile perch and kapenta are fished from Lake Tanganyika.
MT output 1-best	There is a **silver thread and candle from Tanzeri**.
MT hypotheses MC Dropout	There will be a **silver thread and a penny from Tanzer**.
	There is an **attempt at a silver greed and a carpenter from Tanzeri**.
	There will be a **silver bullet and a candle from Tanzer**.
	The **puzzle** is being caught in the **chicken's gavel and the coffin**.
Hypotheses N-best	There is a silver thread and candle from Tanzeri.
	There is a silver thread and candle from Tanzeri.
	There is a silver thread and candle from Tanzeri.

MLQE Dataset

- 7 Language Pairs
- Wikipedia domain
- Manual quality annotation
- 10K sentence pairs per language pair
- NMT systems: SOTA Transformers
- **NMT systems used to generate the translations are available**

https://github.com/facebookresearch/mlqe

https://github.com/sheffieldnlp/mlqe-pe

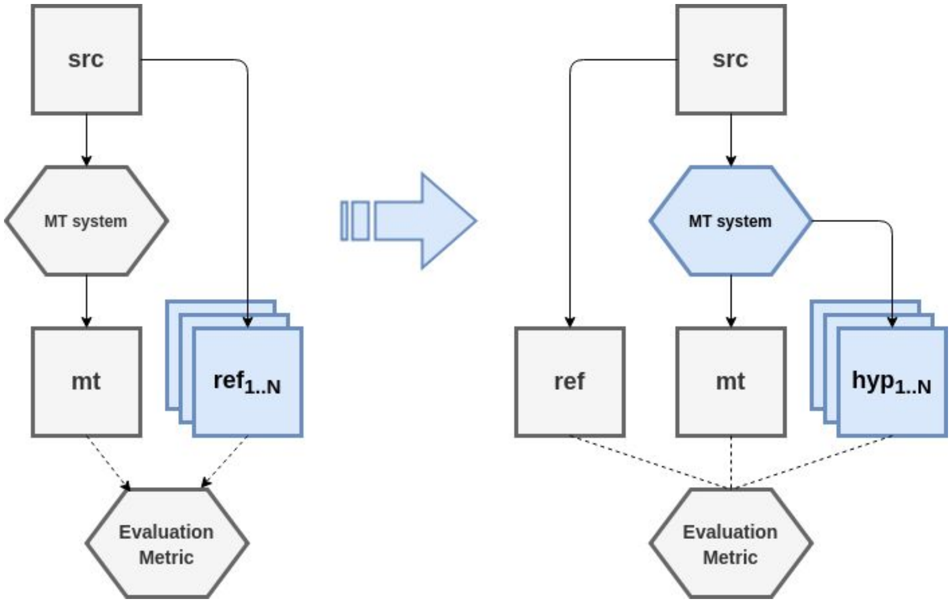
http://www.statmt.org/wmt20/quality-estimation-task.html

The University Of Sheffield.

FB

Unbabel

Glass-box Reference-based Evaluation



- BLEU
- Meteor
- BERTscore
- ...

Multiple references improve MT evaluation

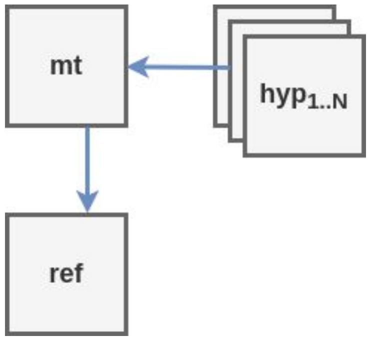
But they are expensive to collect

Use MT hypotheses generated with MC dropout instead

Multi-Hypothesis MT Evaluation

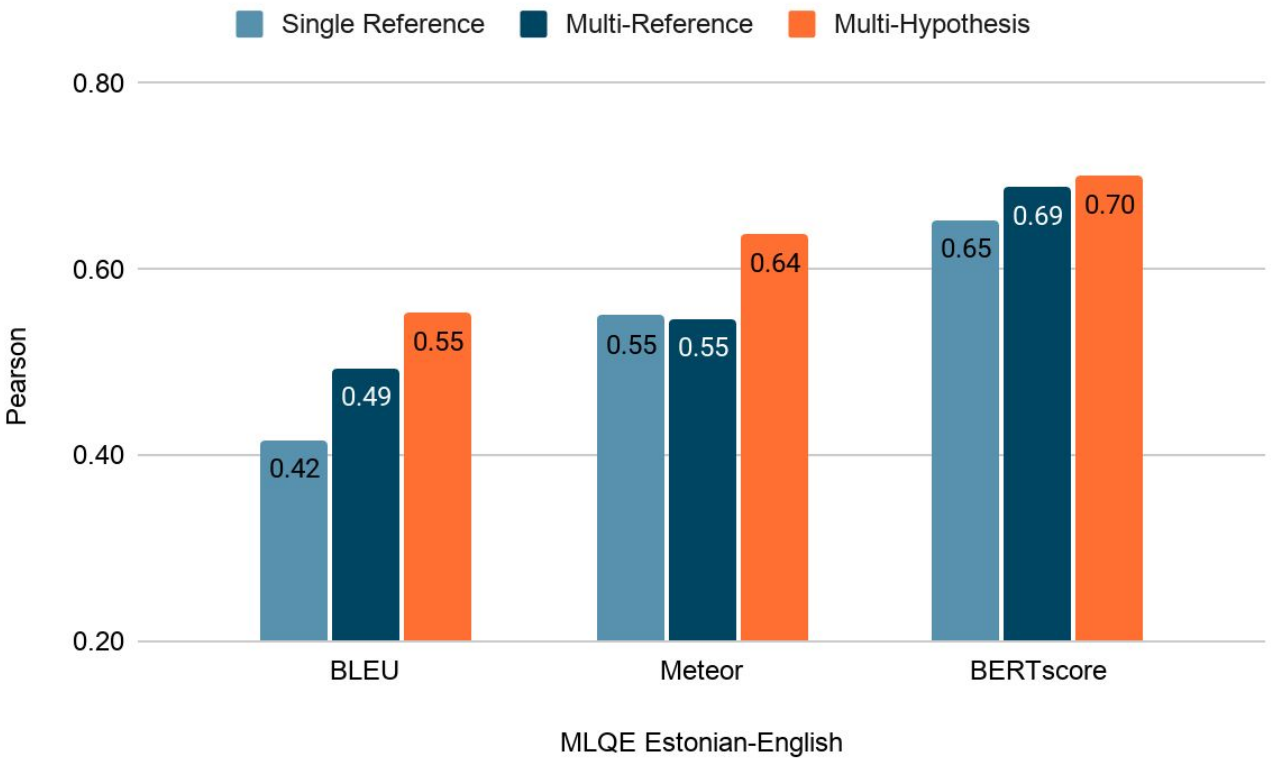
- How to combine this information?

$$\text{hyp-to-mt} = \frac{N^{-1} \sum_{i=1}^{N} sim(hyp_i, mt) + sim(mt, ref)}{2}$$



- Why would this work?
  - Better cover the space of possible solutions
  - Capture predictive uncertainty

Results for Reference-based Evaluation



Glass-box Quality Estimation

1. Unsupervised approach
   - Use attention-based or probability-based metrics directly as quality indicators
2. Lightweight feature-based regression model
   - Train a simple regression model using the indicators as features

Results for Quality Estimation

Method	Low-resource		Mid-resource		High-resource	
	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

Q & A

MLQE dataset
Pearson correlation with human judgements

Results for Quality Estimation

		Low-resource		Mid-resource		High-resource	
	Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
Glass-box	TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
	Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
	D-TP	0.460	0.558	0.642	0.693	0.259	0.321
	D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
	GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
Black-box	PredEst	0.374	0.386	0.477	0.685	0.145	0.190
	Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

MLQE dataset
Pearson correlation with human judgements

Results for Quality Estimation

Method	Low-resource		Mid-resource		High-resource		
	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh	
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*	Unsupervised
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112	
D-TP	0.460	0.558	0.642	0.693	0.259	0.321	
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313	
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**	Supervised
PredEst	0.374	0.386	0.477	0.685	0.145	0.190	
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**	

MLQE dataset
Pearson correlation with human judgements

Results for Quality Estimation

Method	Low-resource Si-En	Low-resource Ne-En	Mid-resource Et-En	Mid-resource Ro-En	High-resource En-De	High-resource En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

TP: Log-probability of MT output

Att-Ent: Entropy of attention weights

Results for Quality Estimation

Method	Low-resource		Mid-resource		High-resource	
	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

D-TP: Average log-probability over *K* forward passes with test-time dropout

D-Lex-Sim: Lexical similarity between *K* hypotheses with test-time dropout

Results for Quality Estimation

Method	Low-resource		Mid-resource		High-resource	
	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

GB-combo: Combination of above indicators as features in a regression model

Results for Quality Estimation

	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

WMT2020 Shared Task on QE

PredEst: Neural-based Predictor-Estimator model [Kim et al., 2017]

Bergamot-LATTE: pretrained contextualized multilingual representations [Sun et al., 2020]

Results for Quality Estimation

Method	Low-resource		Mid-resource		High-resource	
	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

This is better than reference-based evaluation

Results for Quality Estimation

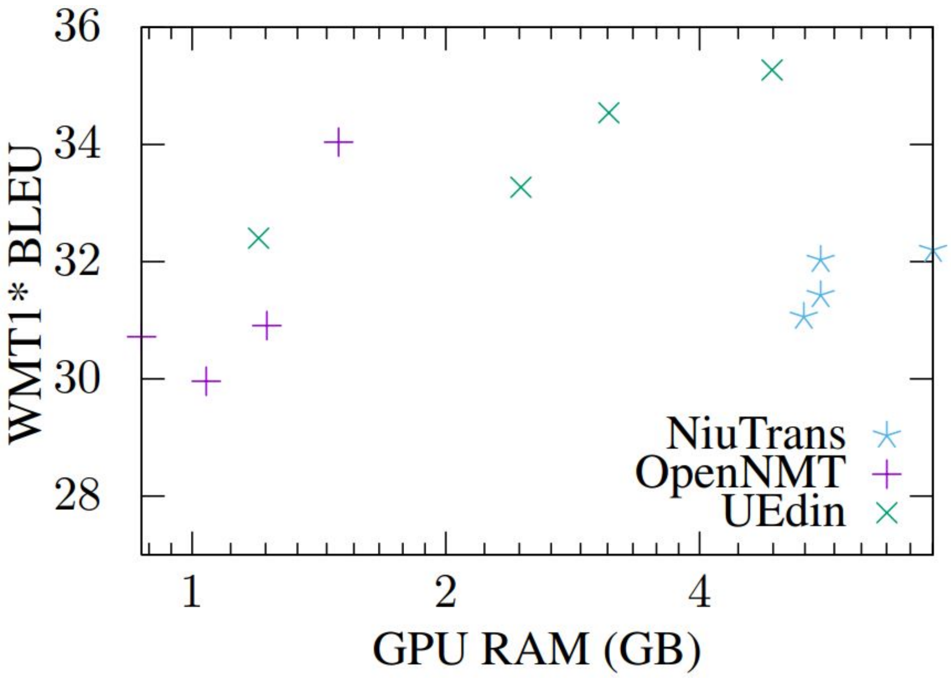
	Low-resource		Mid-resource		High-resource	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

Size of the models

Bergamot-LATTE: **>>561M parameters (> 3G on disk and >6GB in RAM)**

GB-combo: **103 features**

Accuracy-efficiency trade-off



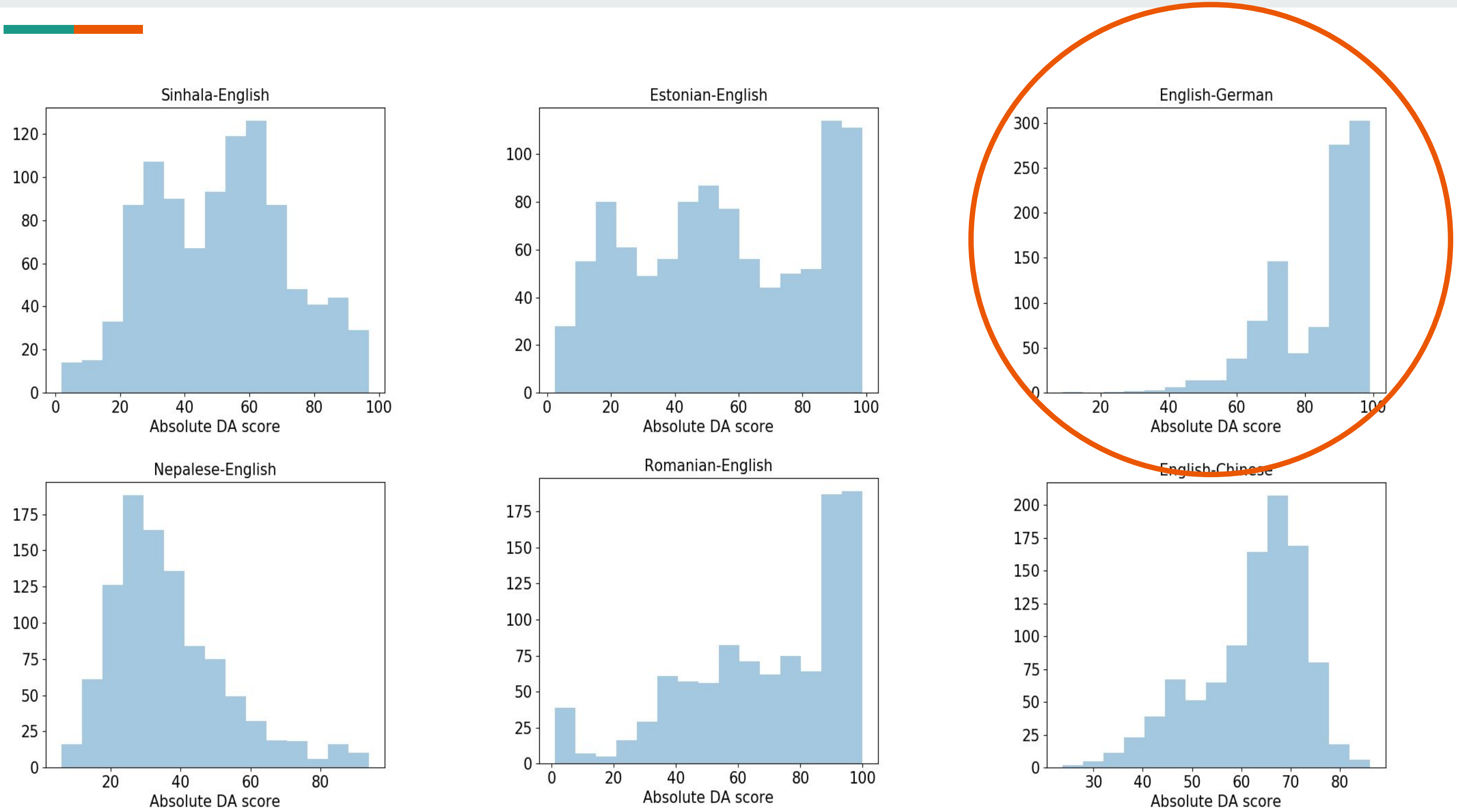
Heafield et al. (2020). Findings of the Fourth Workshop on Neural Generation and Translation

Results for Quality Estimation

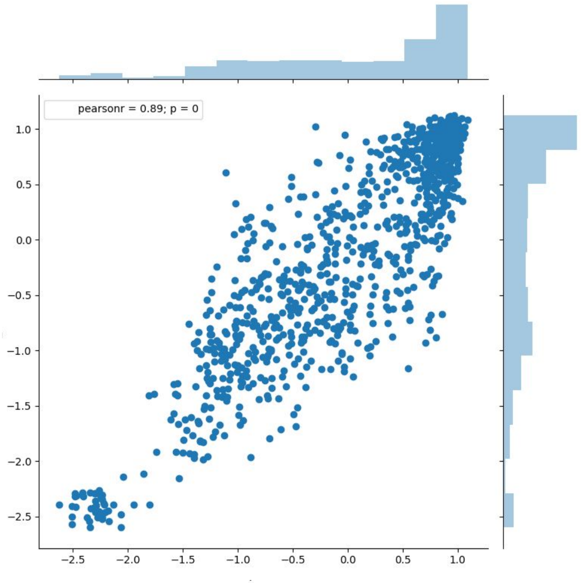
Method	Low-resource Si-En	Low-resource Ne-En	Mid-resource Et-En	Mid-resource Ro-En	High-resource En-De	High-resource En-Zh
TP	*0.399*	*0.482*	*0.486*	*0.647*	*0.208*	*0.257*
Att-Ent (-)	0.100	0.205	0.377	0.382	0.090	0.112
D-TP	0.460	0.558	0.642	0.693	0.259	0.321
D-Lex-Sim	0.513	0.600	0.612	0.669	0.172	0.313
GB-combo	**0.560**	**0.662**	**0.681**	**0.796**	**0.476**	**0.429**
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
Bergamot-LATTE	**0.682**	**0.814**	**0.826**	**0.906**	**0.544**	**0.530**

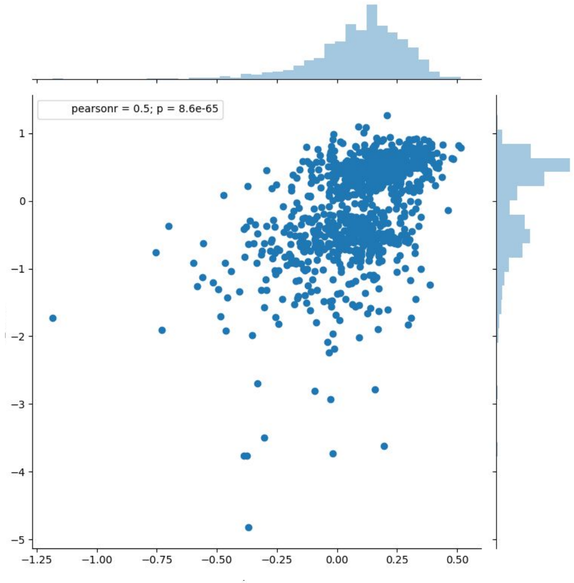
What is wrong with the results for high-resources language pairs?

Distribution of human scores



Distribution of human scores

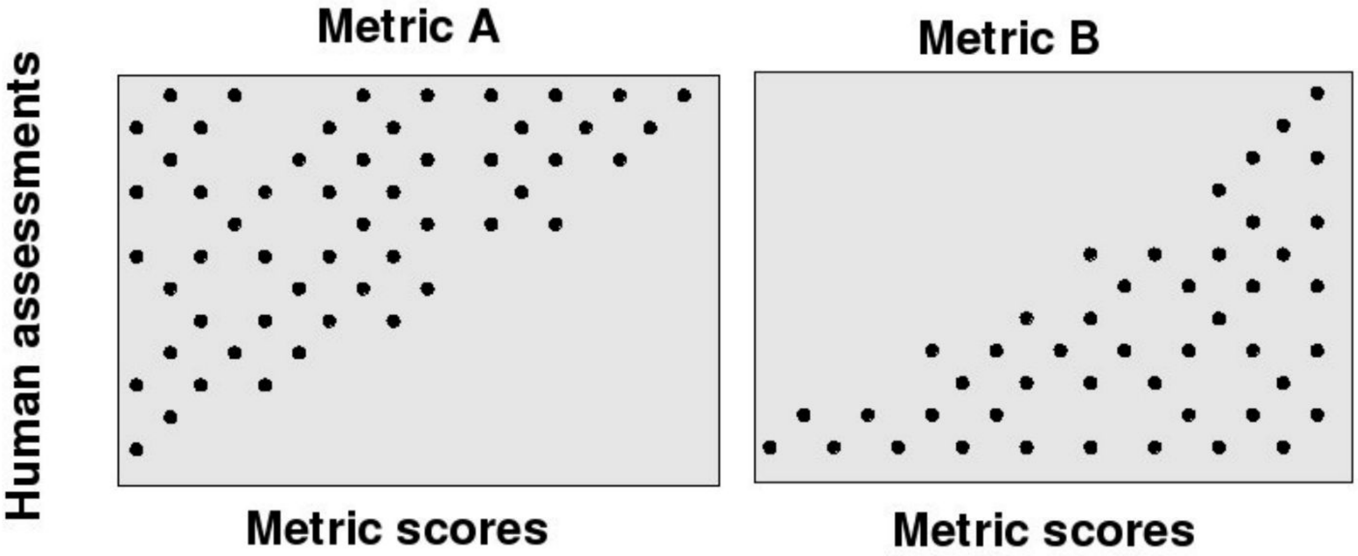




Evaluation of MT Evaluation

MT Evaluation beyond Correlation

Correlation can hide very different behaviours



Amigó et al. (2009)

MT Evaluation beyond Correlation Fomicheva and Specia, 2019

Meta-evaluation study of the behavior of a wide range of reference-based evaluation metrics

What is more challenging to evaluate: low or high-quality MT?

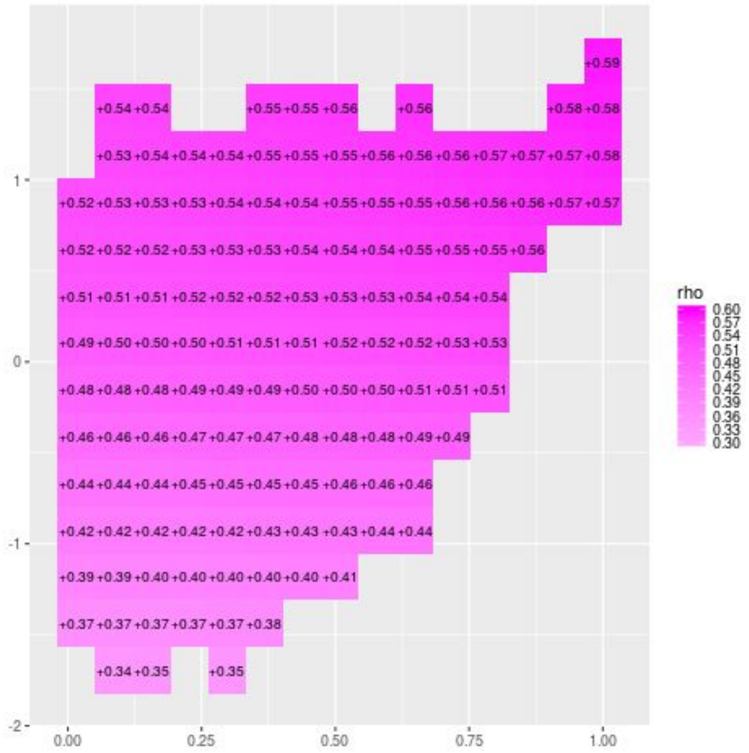
MT Evaluation beyond Correlation

Correlation "breakdown": measure ordinary Pearson correlation in various sub-samples of data

	Q_{low}	Q_{high}	Q^*_{high}	*All*
Meteor	**0.313**	**0.514**†	0.420†	0.570
-TERp-A	0.265	0.459†	0.394†	0.570
MPEDA	**0.313**	0.512†	0.417†	0.568
ROUGE-SU*	0.274	0.453†	0.373†	0.551
ChrF3	**0.321**	0.425†	0.336	0.541
NIST-4	0.258	0.415†	0.327	0.508
BLEU-4	0.159	0.462†	0.360†	0.488
-TER	0.129	0.433†	0.358†	0.462
-WER	0.090	0.458†	0.387†	0.456
-PER	0.175	0.361†	0.281†	0.422

MT Evaluation beyond Correlation

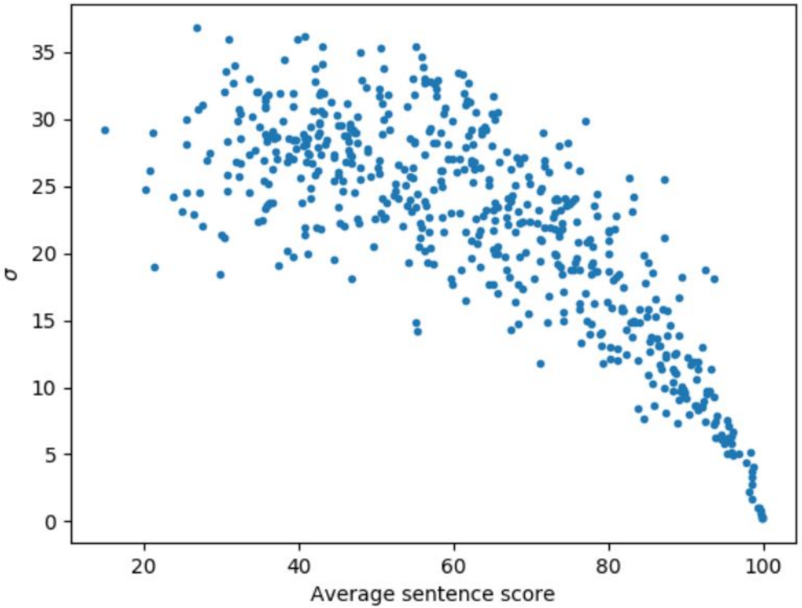
- Correlation “breakdown” can be biased
- Local Gaussian correlation: Fit a gaussian density in the vicinity of each data point
- Confirms that low-quality MT is more challenging for reference-based metrics



https://cran.rstudio.com/web/packages/localgauss/index.html

MT Evaluation beyond Correlation

- Same observation for manual evaluation
- Plot average quality score against the standard deviation of scores assigned to the same sentence by different human judges
- Variability in sentence scores reflects the uncertainty involved in the evaluation process
- Higher variability indicates that the sentence is more difficult to assess



MT Evaluation beyond Correlation

- Possible explanations
  - Low-quality MT outputs contain a higher number of errors
  - For reference-based evaluation metrics
    - Metrics do not measure error severity
    - Lack of informative matches with the reference
  - For humans
    - Perceived impact of different translation errors on the overall translation quality can vary greatly among annotators

Conclusions

Conclusions

- Reference-based and reference-free evaluation should join forces
- Look inside the MT systems (NLG systems?) for useful information
- Quality estimation methods can benefit from all the work on calibration and uncertainty estimation for neural networks
- Pay attention to other aspects beyond correlation with gold labels
  - Properties of the gold label data (distribution, noise, etc.)
  - Model failure modes
  - Accuracy-efficiency trade-off

Think Inside the Box: Glass-box Evaluation Methods for Neural MT

Data: https://github.com/facebookresearch/mlqe

Code:

https://github.com/pytorch/fairseq/tree/master/examples/unsupervised_quality_estimation