cushLEPOR uses LABSE distilled knowledge to improve correlation with human translation evaluations

Gleb Erofeev*, Irina Sorokina*, Lifeng Han^, Serge Gladkoff*

MT Summit 2021

* Logrus Global

^ ADAPT Centre, Dublin City University




The setting (data), and the metrics.
How to measure quality of MT engine candidate?

(And how can we obtain reference evaluation for reference-based metrics?)

Typical Data: TMs

Source	MT Proposal	TM Reference	Reference evaluation	Automated metrics
Lorem ipsum dolor..	...	HQ translation		
Ut enim ad minim..	...	HQ translation		
Duis aute irure dolor ..	...	HQ translation		

BLEU is grossly inaccurate, but readily available for free, e.g. in NLTK
Not much else is available for free
Human evaluations: costly, low agreement, may be biased, and mostly unavailable.
LABSE similarity is excellent proximity measure, but it is difficult to apply and computational-heavy
…we need accurate, simple, fast, free and easily available metrics… customise hLEPOR metric?

BLEU served well - now we need better tool

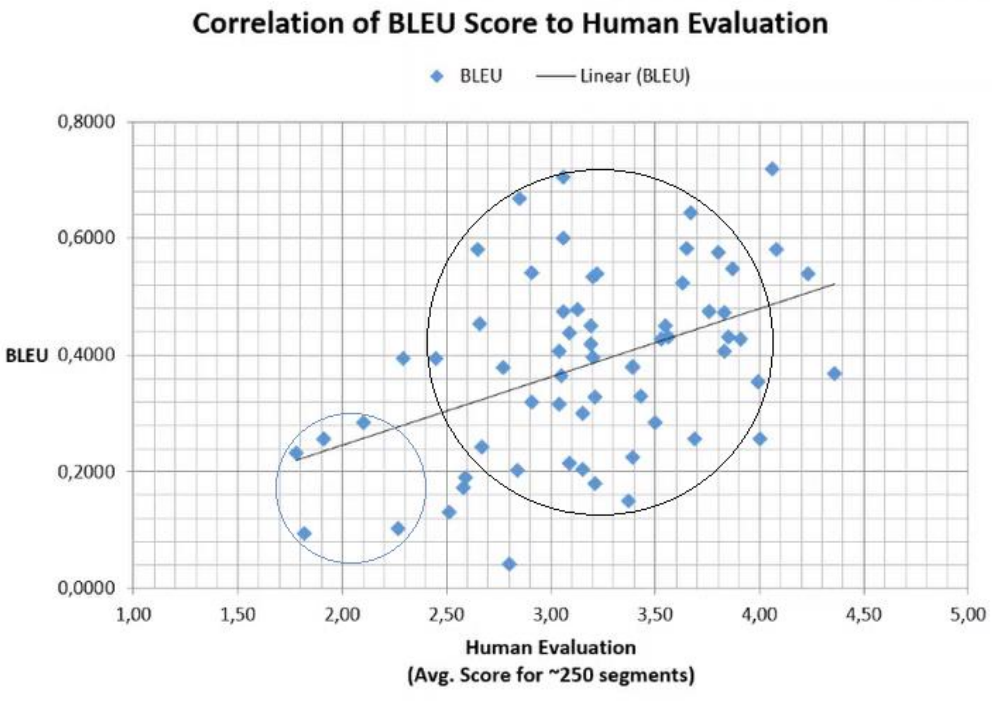
- Very rough measure.
- Inconsistent between implementations.
- Precision-only measure.
- Poor correlation with human judgment

(Was it used most often only because it was readily available for free in nltk?)

Little correlation
with human judgment

A leap of imagination is required to draw a line here, a circle looks much more representative of this scatter.

(c) Diagram courtesy of Jay Marciano, Lengoo



Sample: test set (outside of training set)
Human evaluation: 10% random sampling of test set

Accumulating the pitfalls: ACL2021 outstanding paper award winner

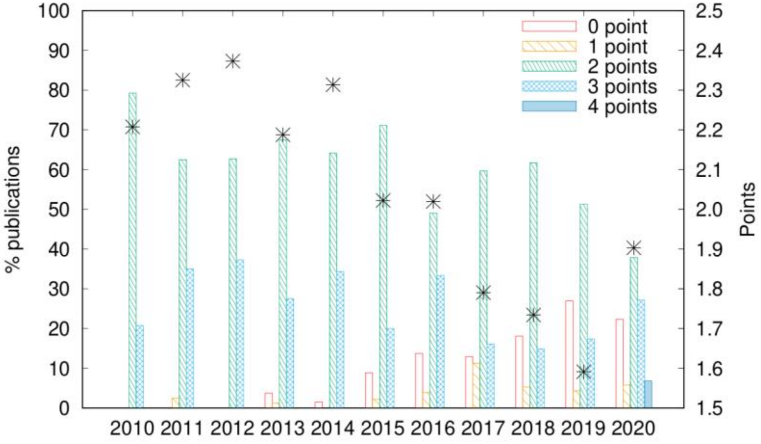
Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

https://aclanthology.org/2021.acl-long.566.pdf

The paper presents the first large-scale metaevaluation of machine translation (MT). “We annotated MT evaluations conducted in 769 research papers published from 2010 to 2020.”

Killer question:

"Is a metric that better correlates with human judgment than BLEU used or is a human evaluation performed?"”



Average mate-eval score (Marie et al. 2021)
MT evaluation worsens.

hLEPOR: best correlation with human judgment

“A Description of Tunable Machine Translation Evaluation Systems in WMT13 Metrics Task” Han et al. 2013: www.statmt.org/wmt13/pdf/WMT53.pdf

hLEPOR includes broader evaluation factors (recall and position difference penalty) in addition to the factors used in BLEU (sentence length, precision), and demonstrated higher accuracy, but Python code was not available.

System	Correlation Score with Human Judgment								
	other-to-English				English-to-other				Mean score
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
LEPOR_v3.1	0.93	0.86	0.88	0.92	0.83	0.82	0.85	0.83	**0.87**
nLEPOR_baseline	0.95	0.61	0.96	0.88	0.68	0.35	0.89	0.83	0.77
METEOR	0.91	0.71	0.88	0.93	0.65	0.30	0.74	0.85	0.75
BLEU	0.88	0.48	0.90	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.50	0.12	0.81	0.84	0.64

hLEPOR (v3.1) on system-level performance using WMT11 data

Directions	EN-FR	EN-DE	EN-ES	EN-CS	EN-RU	Av
LEPOR_v3.1	*.91*	*.94*	*.91*	*.76*	***.77***	***.86***
nLEPOR_baseline	*.92*	*.92*	*.90*	***.82***	*.68*	*.85*
SIMP-BLEU_RECALL	**.95**	.93	.90	**.82**	.63	.84
SIMP-BLEU_PREC	.94	.90	.89	**.82**	.65	.84
NIST-mteval-inter	.91	.83	.84	.79	.68	.81
Meteor	.91	.88	.88	**.82**	.55	.81
BLEU-mteval-inter	.89	.84	.88	.81	.61	.80
BLEU-moses	.90	.82	.88	.80	.62	.80
BLEU-mteval	.90	.82	.87	.80	.62	.80
CDER-moses	.91	.82	.88	.74	.63	.80
NIST-mteval	.91	.79	.83	.78	.68	.79
PER-moses	.88	.65	.88	.76	.62	.76
TER-moses	.91	.73	.78	.70	.61	.75
WER-moses	.92	.69	.77	.70	.61	.74
TerrorCat	.94	**.96**	**.95**	na	na	.95
SEMPOS	na	na	na	.72	na	.72
ACTa	.81	-.47	na	na	na	.17
ACTa5+6	.81	-.47	na	na	na	.17

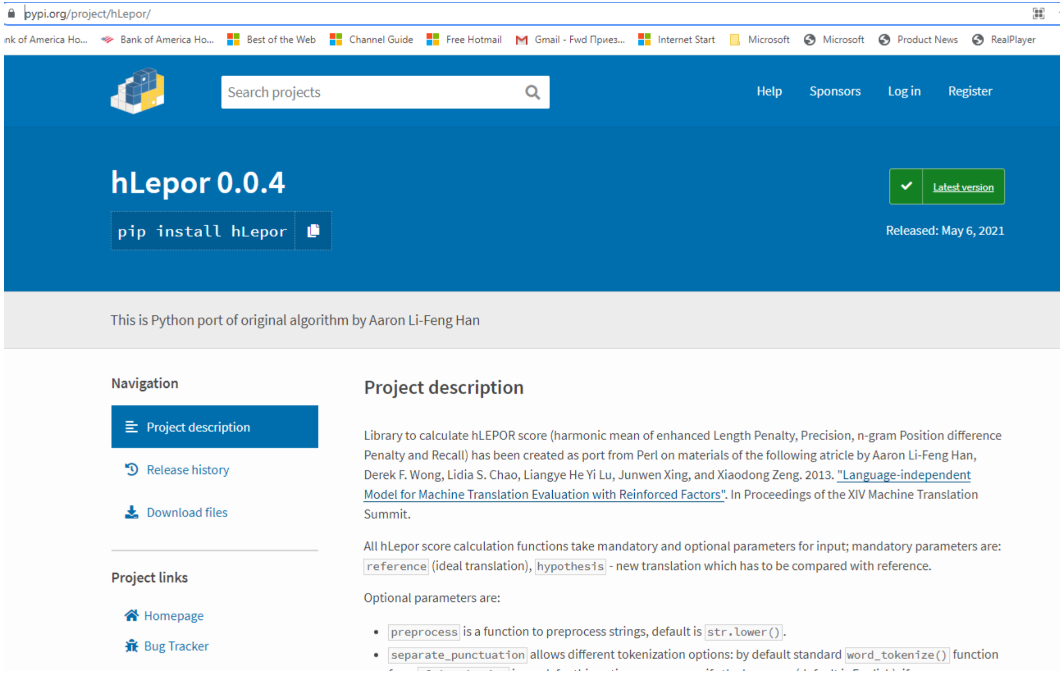
hLEPOR (v3.1) on system-level using WMT13 data, Pearson correlation

under-utilized hLEPOR: we have done Python port:

hLEPOR was ported to Python and published on PyPi.org: https://pypi.org/project/hLepor/

Now it’s available to all engineers and researchers for free!

This version of hLEPOR has 6 customizable parameters!



hLEPOR composition

alpha: the tunable weight for recall
beta: the tunable weight for precision
n: words count before and after matched word in npd calculation
weight_elp: tunable weight of enhanced length penalty
weight_pos: tunable weight of n-gram position difference penalty
weight_pr: tunable weight of harmonic mean of precision and recall

Original hLEPOR takes these parameters as certain suggested empirical values, but how good are they?

Now that we have hLEPOR code, we can try to optimize these parameters against certain data and criteria.

The next step: to fine-tune hLEPOR parameters

In the real world: we don't have human quality evaluations, but we will have TM at best.

How can we get by without the massive involvement of human evaluators, and only engage them for verification of small samples?

One way is to use LABSE similarity measure - Language Agnostic Bert Sentence Embedding by Feng et al. (2020). Its proximity measure shows syntactic similarity very well.

But it is computational-heavy.

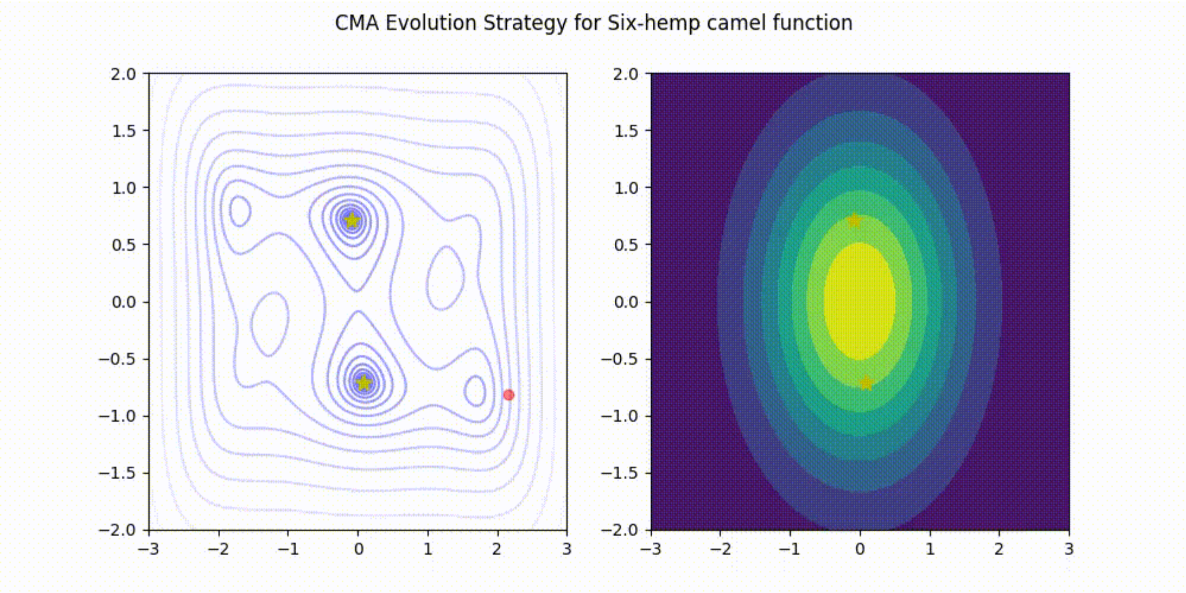
Let’s try to optimize hLEPOR parameters and see if we can improve hLEPOR performance!

(AND we can also try to optimize hLEPOR against human evaluations, too.)

OPTUNA : a hyperparameter optimization network

https://optuna.org/

Optuna is capable of finding the extremums in a seven-dimensional space of 6 parameters and the lowest RMSE (Root Mean Square Error) value.



Left, the optimal solutions (yellow stars) and the solutions sampled by CMA-ES (red points); Right, the update process of the multivariate gaussian distribution.

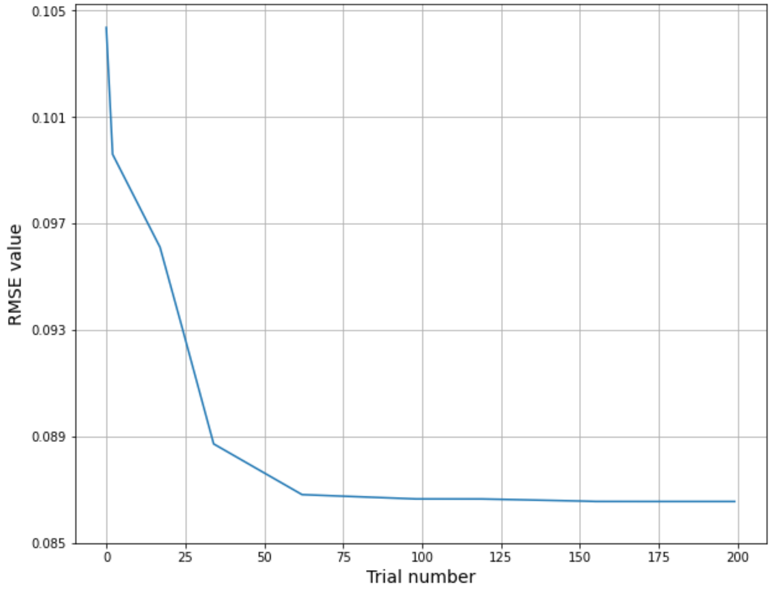
(c) Image courtesy of Masashi SHIBATA

cushLEPOR: customized hLEPOR

1. We build LABSE similarity score on our data.

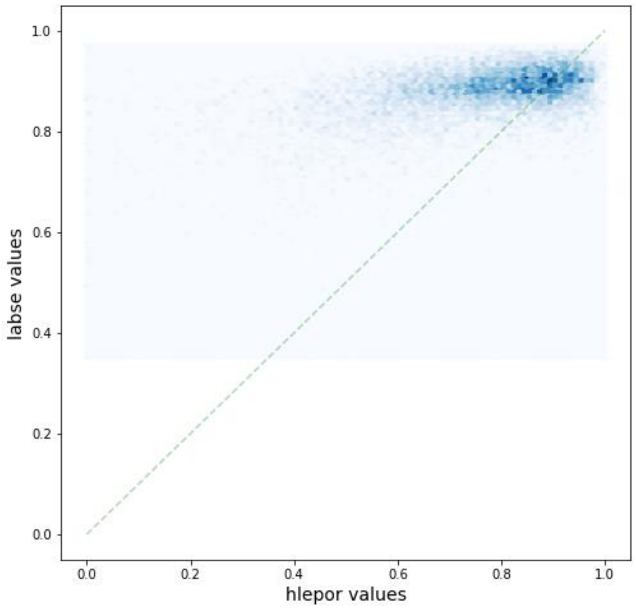
2. We use OPTUNA (https://optuna.org/, a hyperparameter optimization network) to get the lowest possible RMSE (Root Mean Square Error) between cushLEPOR and LABSE

3. The data is available on GitHub: https://github.com/poethan/cushLEPOR

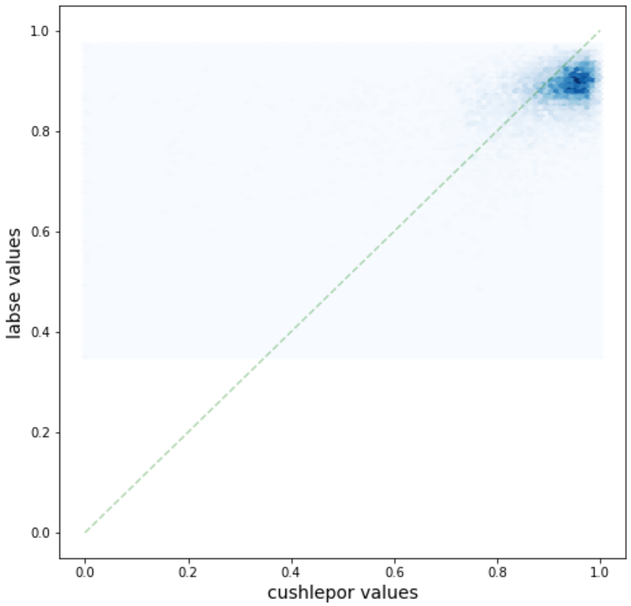


cushLEPOR now shows much better result

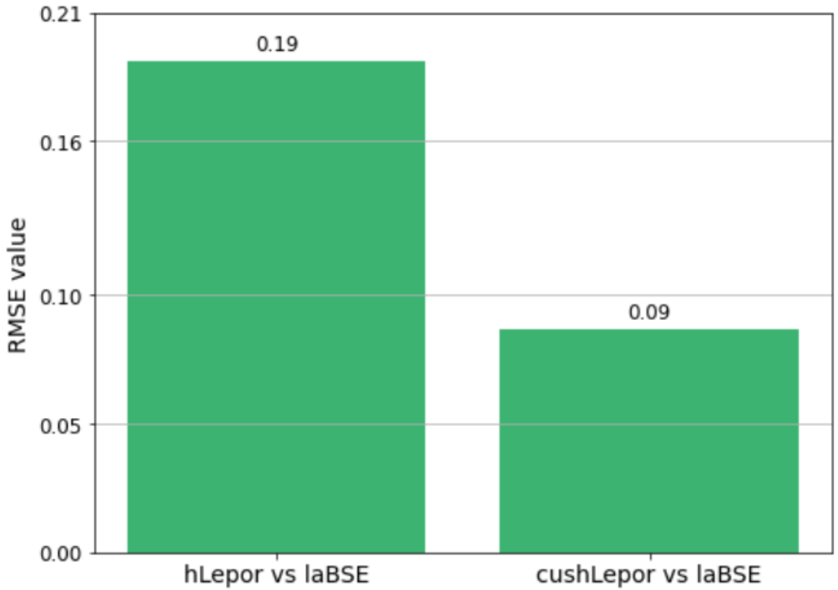
Before:



After:



cushLEPOR(LABSE) has better RMSE than hLEPOR



We have also tried to optimize cushLEPOR vs pSQM

WMT21 shared Metrics tasks suggest using Google Research experiment (with human translator annotated date using MQM and sPQM) for training.

“**Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation**” by Marcus Freitag et.al. (2021) from Google Research: https://arxiv.org/abs/2104.14478

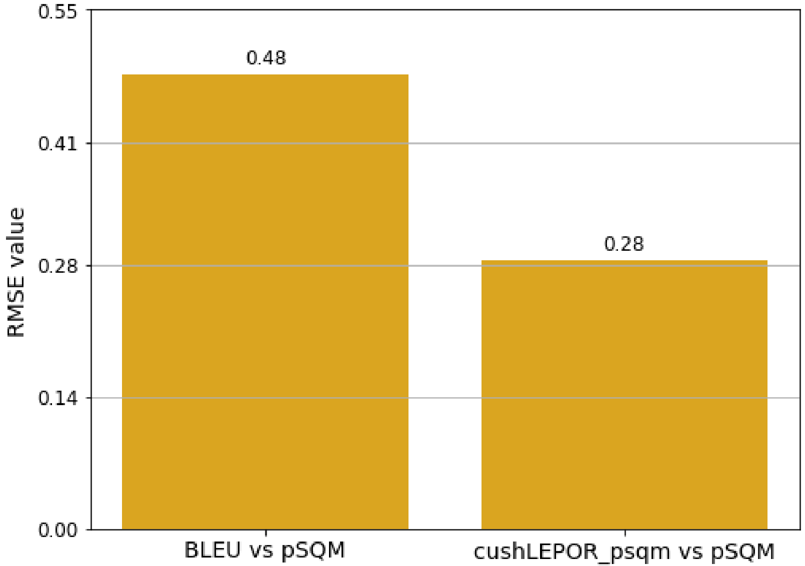
pSQM: professional translator annotated Scalar Quality Metrics
MQM: Multidimensional Quality Metrics (framework)

Features significant corpus of human annotated data with MQM and pSQM metrics.

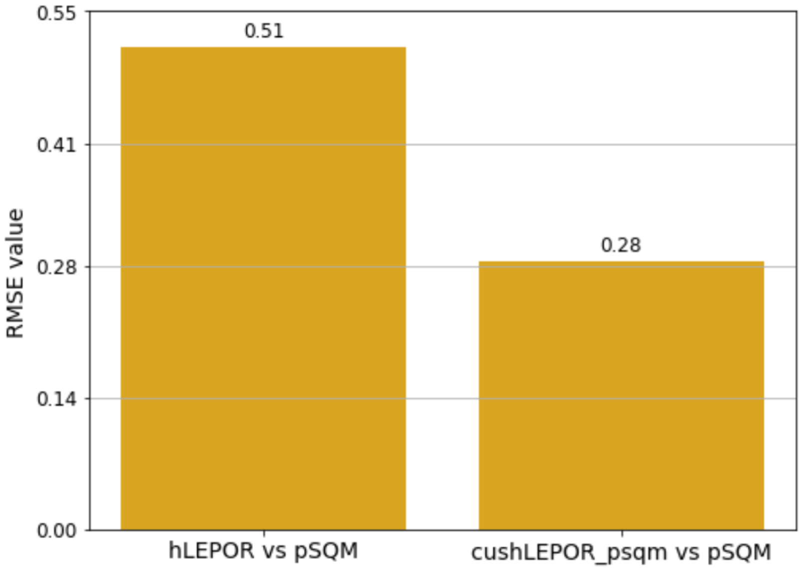
Provides much better results for human judgment.

We have carried out cushLEPOR optimization against MQM and pSQM on En-De and Zh-En.

cushLEPOR(pSQM) gives better RMSE than BLEU



cushLEPOR(pSQM) performs better hLEPOR on pSQM



Conclusions: Advantages

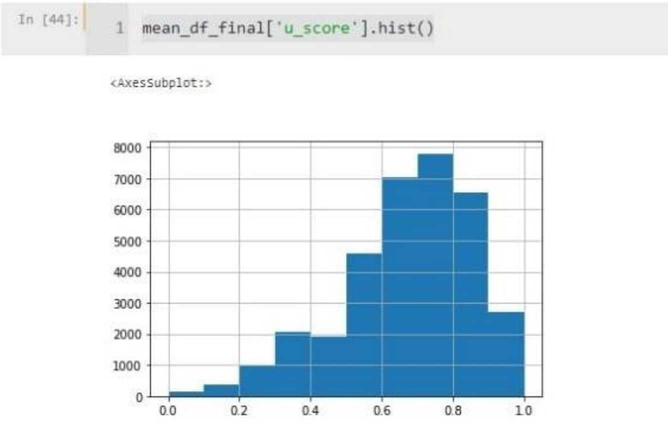
- We now can use cushLEPOR for **target languages** as a light and fast similarity metrics.
- The same code that we have published on PyPi.org can be fine-tuned as cushLEPOR for your application.
- cushLEPOR can be trained on both human evaluations and LABSE similarity.
- N-gram metrics are sensitive to translation variants, but not cushLEPOR because it is optimized for correlation with LABSE (which takes many similar sentences into account as training data).
- LABSE transformer requires IT and ML skills and is computational-heavy. cushLEPOR is an instant light metric that produces the same result after similarity optimization for LABSE.
- Nice simplification of a very complex method.
- cushLEPOR better correlates with human judgment than BLEU, even without our optimization on them.

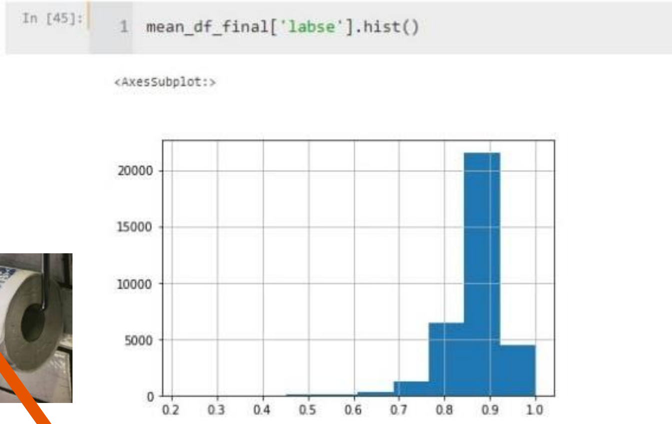
Conclusions: Drawbacks

LABSE and LABSE-optimized cushLEPOR undervalues the significance of errors, error types, showing grammatical syntactic similarity, instead of semantics. Top chart: pSQM human quality ratings distribution. Buttom chart: LABSE similarity measure distribution.

Future work will include semantic features.

In other words, small (from the post-editing point of view) errors may be significant from human perception, but cannot be captured automatically just yet. We plan to analyze different types of errors and assign them different significance (weights) during evaluations.





Conclusions: Practical outcome

You now can use cushLEPOR in actual product.

Do you want us to help you to train your own cushLEPOR for your data and your language pair?

You are welcome.

QUESTIONS?

rd@logrusglobal.com