Poor Man's Quality Estimation

Predicting Reference-Based MT Metrics Without the Reference

What's QE?

Ich esse gerne Pizza mit Ananas.

I eat happy pizza with pineapples.

The issue?

Humans

costly irreproducible low-resource langs.

How do we train QE?

source hypothesis reference

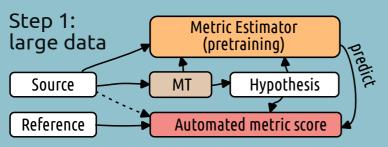
|| qe(s, t) - hum(s, r, t) ||²
human annotation mean squared error

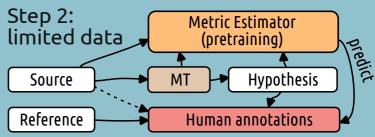
The alternative?

metric(s, r, t) instead of hum(s, r, t)

 $|| qe(s, t) - metric(s, r, t) ||^2$

pretraining on automated metrics, finetuning on humans





Are metrics predictable?

Yes.

ChrF

BLEU

ChrF

METEOR

COMET

TER

Human

TER

(%) uoite au 20

COMET-QE LR TF-IDF LR Multi

ME all ME text

Is this pretraining useful?

A little bit.

Training on human data only: 19.8% Pretraining on TER and finetuning on humans: 22.8%

What else is in the paper?

Transferability of QE across MTs

Synthetic data from beam search

Analysis of what contributes to metric predictability

What's next?

Going beyond MT Complexity & fluency estimation





Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, Mrinmaya Sachan **contact us! vzouhar@inf.ethz.ch**