Google Research

A journey of MT research

Why it is crucial to understand evaluation

Markus Freitag
MT Marathon 2022

Talk

Model	Automatic Eval	Human Eval	Automatic Eval	Model
APE at Scale	BLEU might be Guilty/ WMT20 Metric Task	Expert-based Human Eval	WMT21 Metric Task	Natural Diet/ MBR-Decoding
Goal: Improve naturalness of the MT output	**Goal**: Unbiased automatic metrics	**Goal:** Reliable, explainable human evaluation	**Goal:** Re-evaluate with better ground truth	**Goal:** Find interesting novel translation approaches
Outcome: Automatic Metrics are biased towards literal, non-natural output	**Outcome**: Human raters prefer easy-to-explain output	**Outcome**: Reliable, explainable human evaluation	**Outcome**: Neural metrics correlate well with updated evaluation protocol	**Outcome**: Improvements only visible with updated evaluation protocol

2019

2022

Briefly

In Detail

Briefly

Should we question the metric?
#1
How we start questioning BLEU

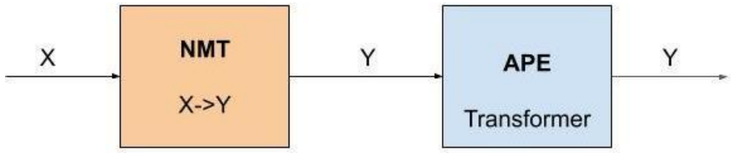
APE at Scale and its Implications on MT Evaluation Biases [WMT19]

Approach

Goal: Improve naturalness of the MT output

2nd goal: Improve accuracy (the APE sees the full translation)

Approach: Automatic post-editing on synthetic data



Outcome

	BLEU	Human
MT	34.3	4.64
MT+APE	30.7	4.63

Automatic Metrics: negative

Crowd Scalar-Value Human Eval: neutral

But - Impression: MT+APE generates more natural, and accurate translations. Problems with evaluation?

What’s wrong with BLEU?

A quick detour

What BLEU actually measures?

BLEU might be Guilty but References are not Innocent

Top matching 4-grams of Facebook with WMT reference:

1. , sagte er → 28 times (, he said .)
2. “ , sagte er → 14 times (” , he said)
3. fuegte hinzu , dass → 8 times (added that)

BLEU might be Guilty but References are not Innocent

Markus Freitag, David Grangier, Isaac Caswell
Google Research
{freitag,grangier,icaswell}@google.com

→ Easy way to generate translation output with high BLEU score:

1. **Match the ngrams responsible for the sentence structure [translating literally gives you the highest success rate - never ever be creative or change the structure!]**
2. **Translate as simple as possible. Using frequent words have a high chance to find a counterpart in the reference translation.**

→ BT, LLMs, MBR Decoding, APE models are typical approaches that improve the output by being more creative and thus yield low BLEU scores.

Should we question the metric?
#2

How we start questioning human evaluation

Translationese as a Language in “Multilingual” NMT + WMT20 Matric Task

Translationese as a Language [ACL 2020]

Goal: Steer Model towards training data with more natural target

Approach: Classify Training data and focus on training examples with natural target

	BLEU	Human Eval
MT	44.6	4.67
NAT MT	41.5	4.72

Similar observation as before. Broken Eval?

WMT20 Metric Task

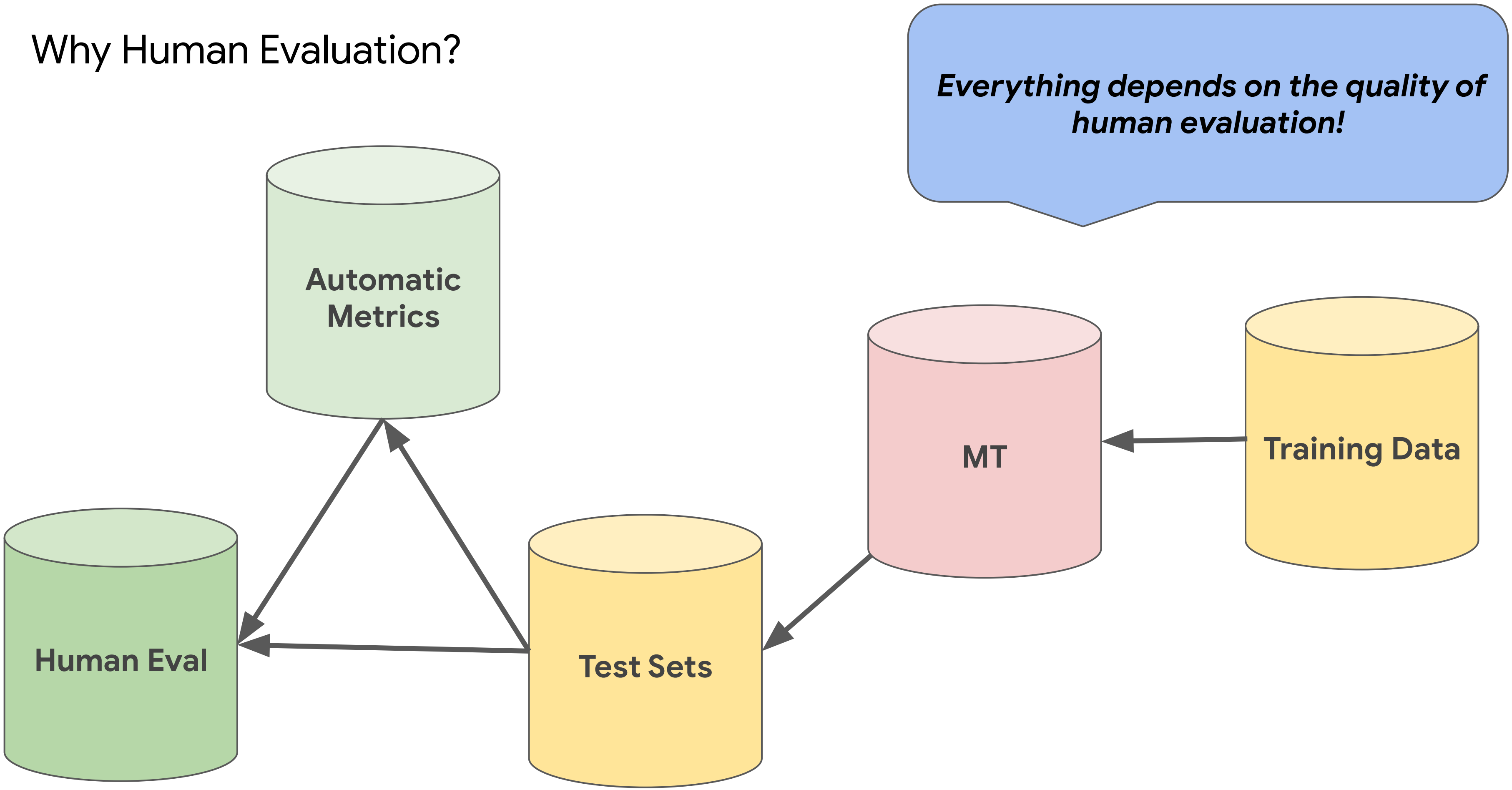
	Rank of human translations
EnDe	1, 4, 10
ZhEn	9

Zh->En	Kendall Tau Correlation
COMET	0.28
BLEURT	0.07

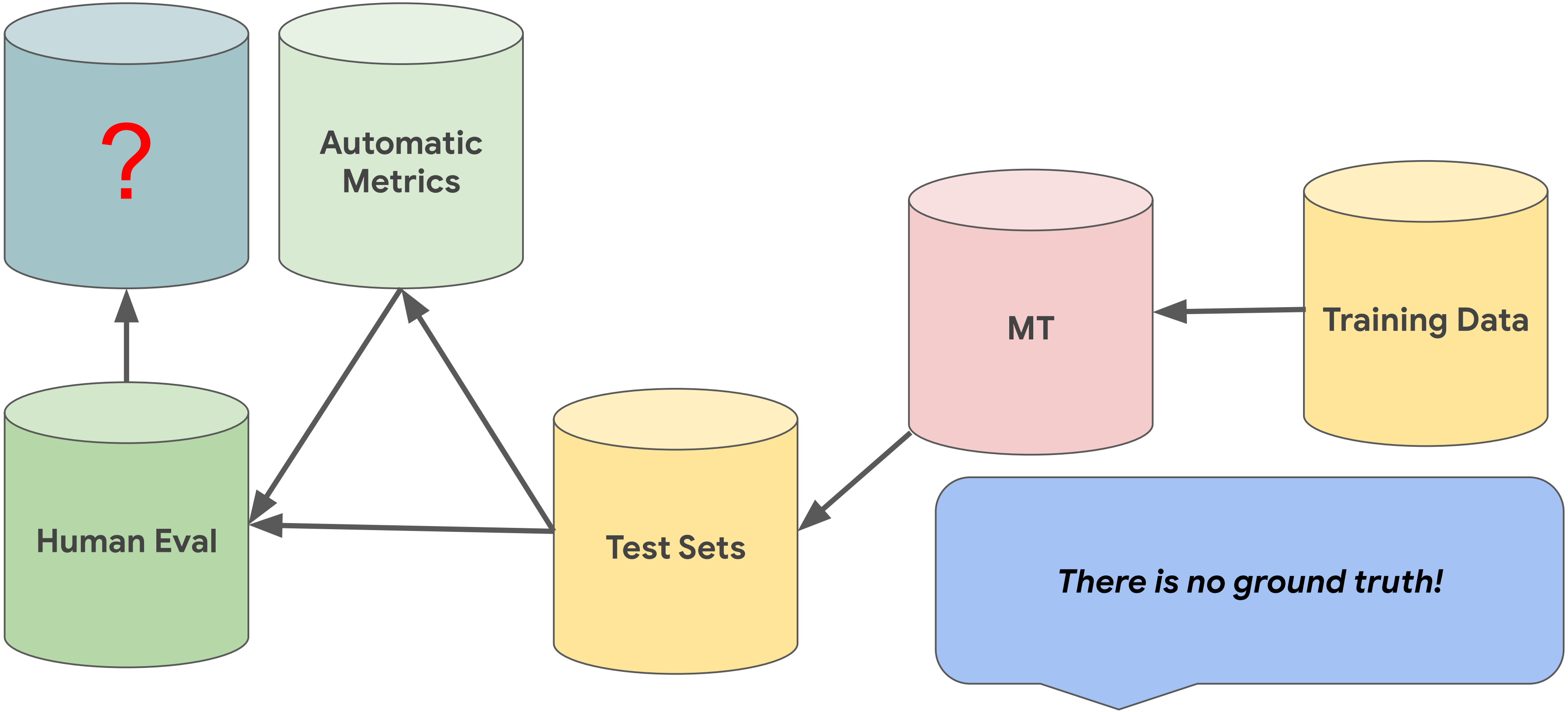
Natural translations, in particular human translations are heavily penalized!

Human Evaluation

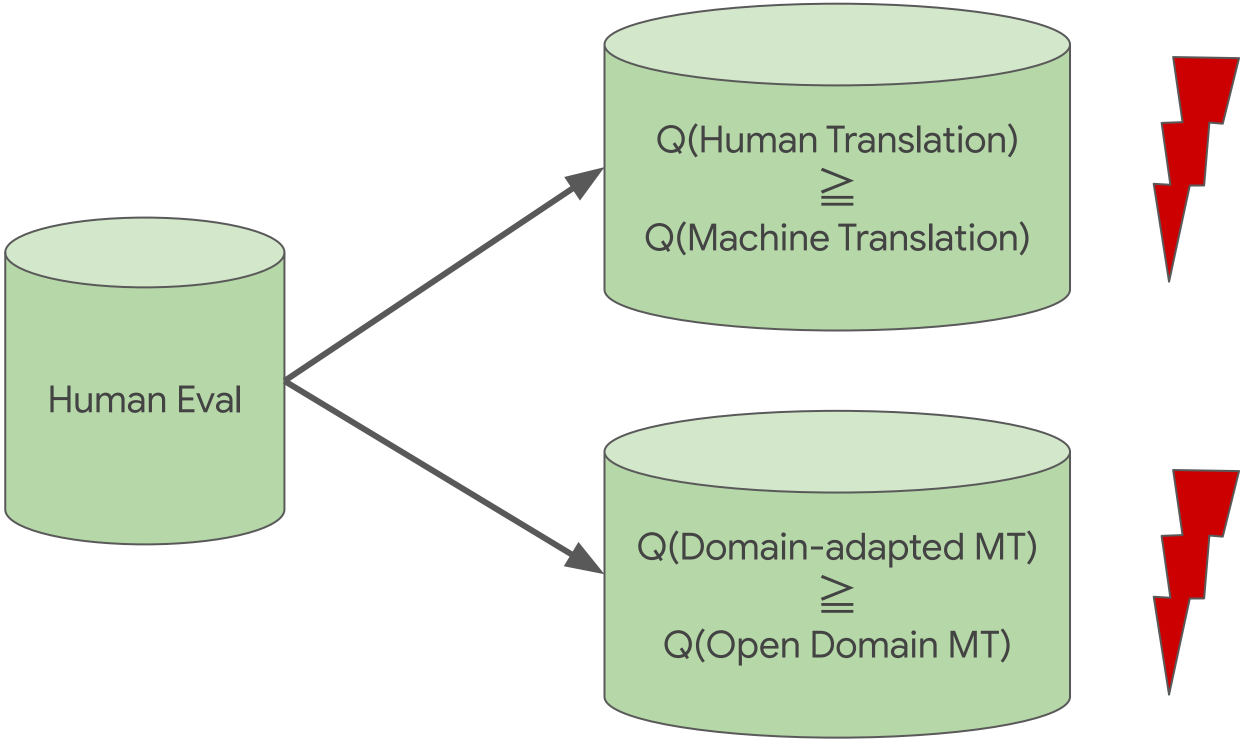
Why Human Evaluation?



How to Measure the Quality of Human Evaluations



Evaluation Based on Anchor Points



Current practices violate these anchor points for high quality language pairs.

Additional Motivation beyond Reliability of System-level scores

Feedback

- Current practices return a scalar value per segment/ system
  - How to interpret the score?
  - Is an improvement real or just rating noise?
- Better feedback:
  - Give details about error categories and severities
  - Is your method really doing what you intended?
  - Helps guide research!
  - Choose error weights for your task

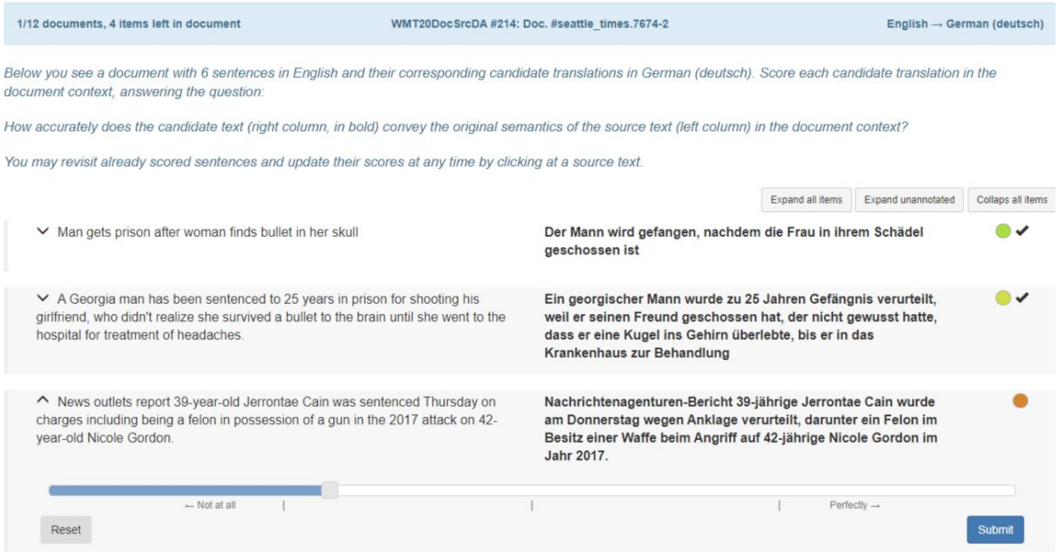
Automatic Metrics

- Automatic metrics heavily rely on segment-level ratings
  - For training and evaluation
- Currently they are noisy and the general impression is that they are very noisy
  - No conclusive answer when comparing 2 metrics
  - Blocker for research in automatic metrics

Current Practices: Example WMT

WMT 2020

- Segment-level ratings with document context (SR+DC) on a 0-100 scale
- Out-of-English:
  - Source-based
  - Rater pool: researchers/translators
- Into-English:
  - Reference-based
  - Rater pool: crowd-workers
- Rater quality control
  - Remove bad ratings
  - Not all segments get a rating
- Z-normalize ratings
  - Put raters on the same scale



Okay, convinced?
What should we do?

Goals of this study:

1. Adapt standards from the (human) translator world

2. Re-evaluate current popular approaches

3. Give recommendations on how to conduct reliable human evaluation

4. Re-evaluate quality of automatic metrics

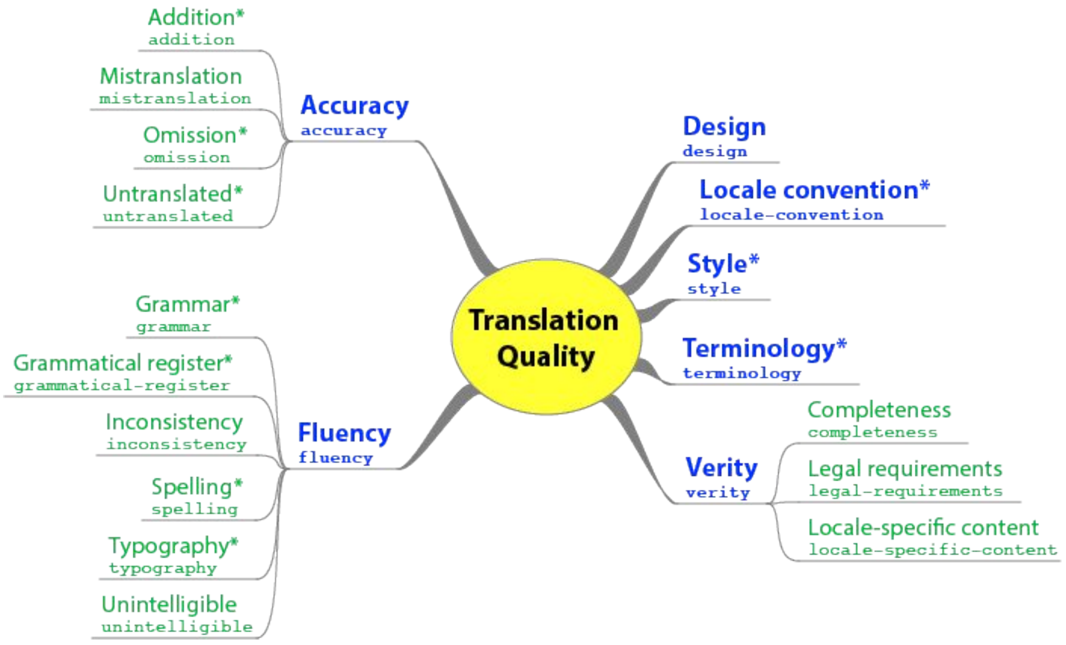
5. Define current error types in machine translation output

Multidimensional Quality Metrics (MQM)

Multidimensional Quality Metrics (MQM)

MQM

- Developed in the EU QTLaunchPad and QT21 projects (www.qt21.eu)
- Generic framework for assessing translation quality, adaptable to various evaluation needs - standard error hierarchy
- Fairly widely adopted by LSPs to evaluate MT and HT. Not so widely adopted by MT researchers.
- To use:
  - Choose relevant errors
  - Choose severity levels
  - Specify weights on errors and severities



MQM Demo

Adapting MQM for Broad-Coverage MT (Translate)

Our MQM schema

- Standard top-level errors: **Accuracy**, **Fluency**, **Terminology**, **Style**, and **Locale** - dedicated sub-categories
- Special error category for completely garbled output: **Non-translation!**
- Three severities:
  - **Major**: real errors
  - **Minor**: imperfections
  - **Neutral**: rater vent
- All standard errors have equal weight except easily-fixable presentation errors

Error Weighting

- Maximum Error count per segment: 25
- System-level error count is the sum of all errors

Severity	Category	Weight
Major	Non-translation	25
	all other	5
Minor	Fluency/Punctuation	0.1
	all other	1
Neutral	all	0

Why is Error Annotation Superior to Asking for a Scalar Value?

Main idea:

- When annotators assign scores or rank translations, their decisions are (or should be?) implicitly based on identifying errors and other imperfections.
- Grounding assessments in explicit error identification creates a "platinum standard" for human evaluation.
    - New, simpler/cheaper schemas can measure correlations to this platinum dataset

Advantages:

- No temptation to "wing it" on long or complex segments
    - Annotators have to "explain" their ratings
    - Fair evaluation of more creative translations!
- Access to annotator rationale, for standardizing ratings and improving systems
- Weight different errors differently depending on the task not on the rater
    - "Taking away the burden of scoring errors"
    - Errors can be differently important for different tasks

Rater Pool - Crowd vs. Prof. Translator

Crowd/Researcher (WMT)

- Pro:
  - Large rater pool
  - Fast evaluation
  - Impact of one rater is minimal
- Cons:
  - Segment-level ratings noisy
  - Needs rater quality control
  - Biases:
    - Prefer the easy, direct translations

Professional Translator (MQM)

- Pro:
  - Native in the target language
  - Fluent in the source language
  - Are trained for the task
  - Reliable segment-level ratings
- Cons:
  - Small rater pool
    - One rater can have a large impact on the final result
  - More expensive
  - Slower turnaround time

Experiments

Experimental Setup - WMT 2020

WMT Submissions

- Top submission of WMT2020
- 6 for ZhEn, 5 for EnDe
- Very similar systems
- Domain-adapted systems
- The best translations we can generate for the news-domain

WHY

- Used for research

Online Systems

- 2 online systems:
- Online-A
- Online-B

WHY

- Different training data
- Not tuned on news-translations
- Worse quality on news domain -> Should be ranked last

Human Translations

- 2 standard human translations (Human-A, Human-B)
- 1 Paraphrased translation for EnDe (Human-P)
- Generated in-context

WHY

- The future of MT
- Should be ranked ahead of MT

English->German System-level Rankings

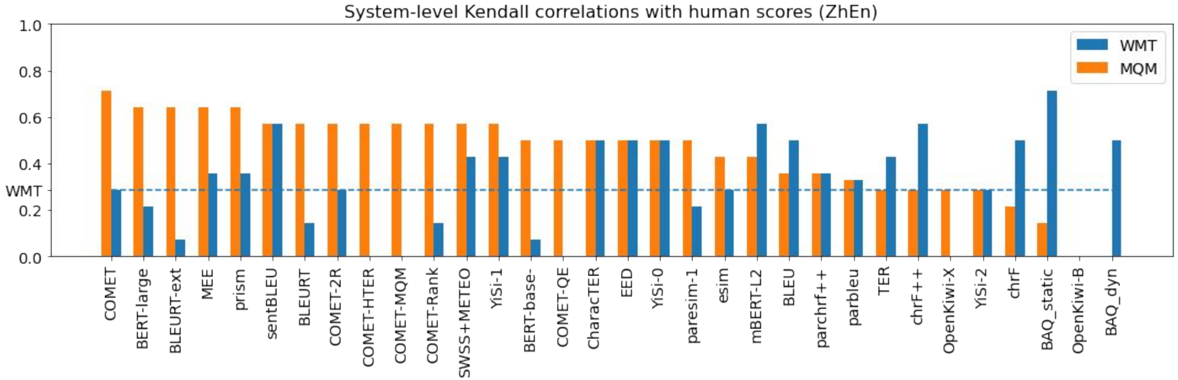
System	WMT↑	MQM↓
Human-B	0.57(1)	0.75(1)
Human-A	0.45(4)	0.91(2)
Human-P	0.30(10)	1.41(3)
Tohoku-AIP-NTT	0.47(3)	2.02(4)
OPPO	0.50(2)	2.25(5)
eTranslation	0.31(9)	2.33(6)
Tencent_Translation	0.39(6)	2.35(7)
Huoshan_Translate	0.33(7)	2.45(8)
Online-B	0.42(5)	2.48(9)
Online-A	0.32(8)	2.99(10)

- MQM correctly ranks the anchor points
  - Human translations are ranked higher than MT
  - Human-P are paraphrased translations - very challenging to evaluate
- WMT has low correlation with MQM
  - Revising the system ranking in WMT20

Impact on Automatic Evaluation: Metrics from WMT20 Task

Zh->En	WMT ratings	MQM ratings
COMET	0.28	0.71
BLEURT	0.07	0.64
WMT	n/a	0.28

System-level Kendall tau



- Correlation of metrics is very different when comparing to MQM (orange) vs WMT (blue).
- Dotted line is MQM/WMT (human/human) correlation.
- Most metrics outperform WMT human scores!

Error Category Distribution

Error Categories	Errors (%)	Major (%)	Human MQM	All MT MQM	All MT vs H.	Tohoku MQM	Tohoku vs H.	OPPO MQM	OPPO vs H.	eTrans MQM	eTrans vs H.
Accuracy/Mistranslation	33.2	27.2	0.296	1.285	*4.3*	**1.026**	**3.5**	1.219	*4.1*	1.244	*4.2*
Style/Awkward	14.6	4.6	0.146	0.299	*2.0*	0.289	*2.0*	0.315	*2.1*	0.296	*2.0*
Fluency/Grammar	10.7	4.7	0.097	0.224	*2.3*	0.193	*2.0*	0.215	*2.2*	0.196	*2.0*
Accuracy/Omission	3.6	13.4	0.070	0.091	*1.3*	0.063	*0.9*	0.063	*0.9*	**0.120**	**1.7**
Accuracy/Addition	1.8	6.7	0.067	0.025	*0.4*	0.018	*0.3*	0.024	*0.4*	0.021	*0.3*
Terminology/Inappropriate	8.3	7.0	0.061	0.193	*3.2*	0.171	*2.8*	0.189	*3.1*	0.193	*3.2*
Fluency/Spelling	2.3	1.2	0.030	0.039	*1.3*	0.030	*1.0*	0.039	*1.3*	0.028	*0.9*
Accuracy/Untranslated text	3.1	14.9	0.024	0.090	*3.8*	0.082	*3.5*	0.066	*2.8*	0.098	*4.2*
Fluency/Punctuation	20.3	0.2	0.014	0.039	*2.8*	**0.067**	**4.9**	0.013	*1.0*	0.011	*0.8*
Other	0.5	5.2	0.005	0.010	*1.9*	0.009	*1.6*	0.010	*1.9*	0.007	*1.2*
Fluency/Register	0.6	5.0	0.005	0.014	*3.0*	0.009	*1.9*	0.015	*3.2*	0.015	*3.3*
Terminology/Inconsistent	0.3	0.0	0.004	0.005	*1.2*	0.004	*0.9*	0.005	*1.2*	0.005	*1.2*
Non-translation	0.2	100.0	0.003	0.083	*28.3*	0.041	*14.0*	0.065	*22.0*	**0.094**	**32.0**
Fluency/Inconsistency	0.1	1.3	0.003	0.002	*0.7*	0.001	*0.3*	0.001	*0.3*	0.003	*1.0*
Fluency/Character enc.	0.1	3.7	0.002	0.001	*0.7*	0.002	*1.0*	0.001	*0.6*	0.000	*0.2*
All accuracy	41.7	24.2	0.457	1.492	*3.3*	1.189	*2.6*	1.372	*3.0*	1.483	*3.2*
All fluency	34.2	1.8	0.150	0.320	*2.1*	0.303	*2.0*	0.284	*1.9*	0.253	*1.7*
All except acc. & fluenc	24.2	6.0	0.222	0.596	*2.7*	0.526	*2.4*	0.591	*2.7*	0.596	*2.7*
All categories	100.0	12.1	0.829	2.408	*2.9*	2.017	*2.4*	2.247	*2.7*	2.332	*2.8*

MQM gives feedback!

1. **Tohoku:**
   Fewer mistranslations
   More punctuation errors
2. **eTrans:**
   More Omission errors
   More Non-Translations!

Impact of Improved Human Evaluation Protocol

WMT21 Metric Task

Expert-based Human Evaluation

Changes:

- Jointly with Unbabel:
    - Expert-based human ratings of WMT submissions with MQM for 3 language pairs
- Addition of interesting metric systems that are challenging for both humans and machines
- Evaluation beyond the mean metric scores!

Goal:

- Establish standard + tooling for both human and automatic eval

WMT21 Results

Results in line with the ones of WMT20

- MQM ranks human translations higher than MT and correlates much better with metrics
- WMT DA scores correlate poorly with MQM and metrics

Use MQM annotation for metric research

- We would like to encourage everyone working on metrics to use the MQM annotation as groundtruth.
- All data is available on github!

A Natural Diet: Towards Improving Naturalness of Machine Translation Output [ACL 22]

An Example of Error-driven Research

Approach

Natural target → Add <nat> tag

Translation target → Add <tran

Shows that method works!
But: Introduces new errors (see major-mistranslation)

based on the nature
training example.

Inference: Add <nat> tag to the source sentence

Experimental Human Results

MQM EnDe (extract)

	base		<nat>	
	Major	Minor	Major	Minor
Mistranslation	44	79	51	26
Grammar	14	56	20	29
.	.	.	.	.
.	.	.	.	.
Awkward	14	143	7	95
Total Errors	113	395	132	275
Global Score	0.91		0.88	

Awkward category significantly improved, while mean score is almost neutral

The impact of better understanding

Minimum Bayes Risk Decoding with Neural Metrics
High Quality Rather than High Model Probability

Motivation: MAP Decode for Translation Task

system	translations	log pplx
source	Der Ausbruch sei „**mit Ansage**“ **gekommen**.	
beam (=4)	The outbreak **came** "**with announcement**".	-2.82
human-A	The outbreak **occurred** “**predictably**”.	-18.10
human-B	The outbreak **happened** “**on cue**.”	-18.74

- Beam search:
  - Generates target words that are frequent in the training data
  - Translates very literal without considering the sentence context too much
- Consequences
  - **Domain mismatch** (generate most likely word-to-word translation)
  - Can introduce **omission** errors (words “hard” to explain)
  - Sounds **awkward and unnatural**

- Human translation have very low pplx as humans use words and sentences structures that are rare in the training data!

Motivation: MAP Decode for Translation Task

system	translations	log pplx
source	Der Ausbruch sei „**mit Ansage**“ **gekommen**.	
beam (=4)	The outbreak **came** "**with announcement**".	-2.82
model sample	The outbreak **happened** "**with announcement**".	-6.03
model sample	The outbreak **occurred** "**with announcement**".	-6.61
model sample	(Table of Contents)	-16.15
model sample	The outbreak took a "say-so".	-18.38
human-A	The outbreak **occurred** “**predictably**”.	-18.10
human-B	The outbreak **happened** “**on cue**.”	-18.74

- Potential solutions:
  - Training data distribution
  - Training objective / model
  - Inference strategy (in this talk)
    - Instead of the most likely translation (based on the model), we generate the most acceptable translation

Minimum Bayes Risk Decoding

Are we using the right **decoding criterion**?

- Beam search = Maximum LogP ≠ Maximum utility (BLEU, BLEURT, YISI, CHRF...)

MBR decoding

- Take S, i.e. N samples from model and find the utility "***centroid***"

$$h_{\mathrm{MBR}} = \arg\max_{h \in S} \frac{1}{N} \sum_{h' \in S}^{N} u(h; h')$$

- Need:
  - good model (probability distribution good estimate of $P_{\mathrm{human}}(y|x)$
  - good utility (BLEURT, YISI, CHRF, BLEU?)

Utility Functions

EnDe newstest2021	log pplx	BLEU	Chrf	YiSi	BLEURT
Human Translation	-38.0	31.5	60.9	84.7	37.1
Beam (=4)	-11.5	35.2	63.0	85.6	30.3

- **BLEU, Chrf and YiSi**
  - Word/emb-based overlap metrics
  - Aligned with log ppl
  - Idea: explain every single token in the hypotheses with tokens in the reference
- **BLEURT**
  - Projects sentences into an embedding space
  - The sentence structure and the actual token play a secondary role
  - The semantic and the fluency are important

Experimental Setup

- **Language Pairs**:
  - De<->En
- **Training data**:
  - WMT2019 - 57M parallel sentences (paracrawl, nc-v15, europarl, commoncrawl)
  - Filtered via CDS (indomain = nc-v15)
- **Test set**:
  - newstest2019 (dev), newstest2021 (test)
- **Model**:
  - Transformer-big, 300k training steps
  - Model trained w/o label smoothing
- **MBR Decoding**:
  - Sampling strategy: ancestral sampling
  - Candidate Size: 1000
  - Utility function: sentenceBLEU, Chrf, YiSi, BLEURT

English⟶German newstest2021 (ref-C)

	BLEU	sBLEU	Chrf	YiSi	BLEURT
Human Translation (ref-D)	31.5	31.6	60.9	84.7	75.6
Beam (=4)	**34.3**	34.2	62.5	85.3	71.6

English→German newstest2021 (ref-C)

	BLEU	sBLEU	Chrf	YiSi	BLEURT
Human Translation (ref-D)	31.5	31.6	60.9	84.7	75.6
Beam (=4)	**34.3**	34.2	62.5	85.3	71.6
MBR-sBLEU-add_k(k=1)	34.7	***34.8***	62.5	85.4	70.5
MBR-CHRF	34.2	34.3	***64.1***	85.7	71.4
MBR-YiSi	34.2	34.2	62.8	***86.0***	71.6

English→German newstest2021 (ref-C)

	BLEU	sBLEU	Chrf	YiSi	BLEURT
Human Translation (ref-D)	31.5	31.6	60.9	84.7	75.6
Beam (=4)	**34.3**	34.2	62.5	85.3	71.6
MBR-sBLEU-add_k(k=1)	34.7	***34.8***	62.5	85.4	70.5
MBR-CHRF	34.2	34.3	***64.1***	85.7	71.4
MBR-YiSi	34.2	34.2	62.8	***86.0***	71.6
MBR-BLEURT	25.4*	26.0	57.7	83.1	***79.0***

*For more details of the biases of BLEU and why it is good to reduce BLEU scores, read:
BLEU might be Guilty but References are not Innocent (EMNLP 2020)

English→German newstest2021 (ref-C)

	BLEU	sBLEU	Chrf	YiSi	BLEURT	log ppl
Human Translation (ref-D)	31.5	31.6	60.9	84.7	75.6	-38.0
Beam (=4)	***34.3***	34.2	62.5	85.3	71.6	-11.5
MBR-sBLEU-add_k(k=1)	34.7	***34.8***	62.5	85.4	70.5	-11.2
MBR-CHRF	34.2	34.3	***64.1***	85.7	71.4	-13.2
MBR-YiSi	34.2	34.2	62.8	***86.0***	71.6	-11.4
MBR-BLEURT	25.4	26.0	57.7	83.1	***79.0***	-24.4

English→German newstest2021 (ref-C)

	BLEU	sBLEU	Chrf	YiSi	BLEURT	log ppl	MQM human eval ↓
Human Translation (ref-D)	31.5	31.6	60.9	84.7	75.6	-38.0	0.338
Beam (=4)	**34.3**	34.2	62.5	85.3	71.6	-11.5	2.392
MBR-sBLEU-add_k(k=1)	34.7	***34.8***	62.5	85.4	70.5	-11.2	**1.992**
MBR-CHRF	34.2	34.3	***64.1***	85.7	71.4	-13.2	**2.214**
MBR-YiSi	34.2	34.2	62.8	***86.0***	71.6	-11.4	2.842
MBR-BLEURT	25.4	26.0	57.7	83.1	***79.0***	-24.4	**1.562**

1. MBR works: **Each MBR-utility is best on their utility**
2. Human evaluation
   a. **MBR-BLEU outperforms beam search decoding**
   b. **MBR-BLEURT wins** by a huge margin

MQM - Human Evaluation with Error Categories

Error Category	beam	MBR BLEURT
Accuracy/Omission	18	7
Terminology/Inappropriate for context	151	89
Style/Awkward	66	46
Accuracy/Mistranslation	70	58

Number of major errors

Improvements:

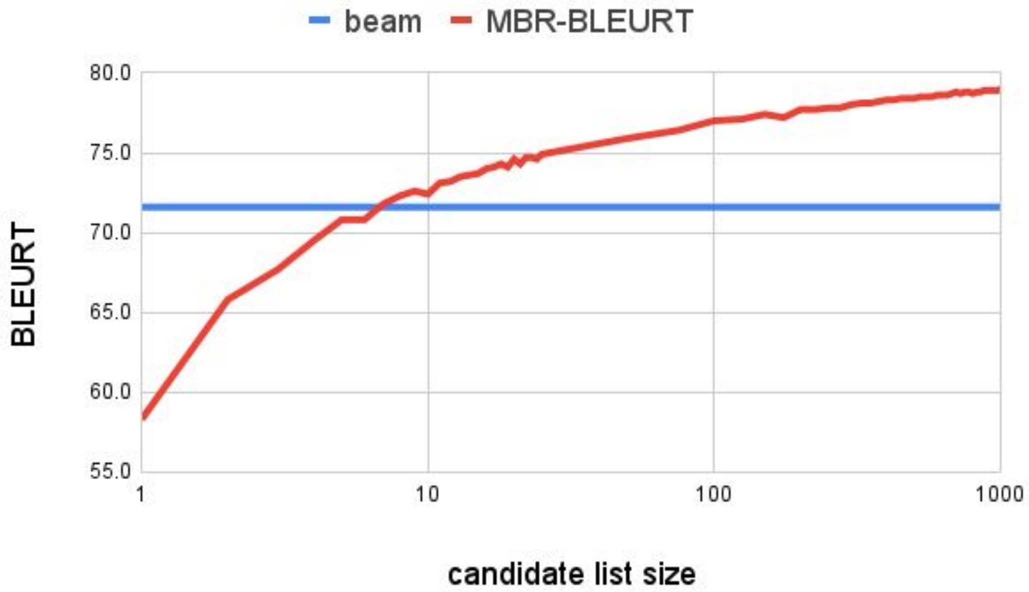
1. Beam Search Problem:
    - Length of translations
    - Example: omission
2. Beam Search Problem:
    - Autoregressive decoding
    - Example: inappropriate for context
3. Beam Search Problem:
    - Generate most probably tokens
    - Example 1: Style/ Awkward
    - Example 2: Mistranslation

Example Translations

System	translations	log pplx
source	Der Ausbruch sei „**mit Ansage**“ **gekommen**.	
beam (=4)	The outbreak **came** "**with announcement**".	-2.82
MBR-sBLEU	The outbreak **came** "**with announcement**".	-2.82
MBR-Chrf	The outbreak **came** "**with announcement**".	-2.82
MBR-YiSi	The outbreak **came** "**with announcement**".	-2.82
MBR-BLEURT	The outbreak **occurred** "**as announced**".	-11.21
human-A	The outbreak **occurred** “**predictably**”.	-18.10
human-B	The outbreak **happened** “**on cue**.”	-18.74

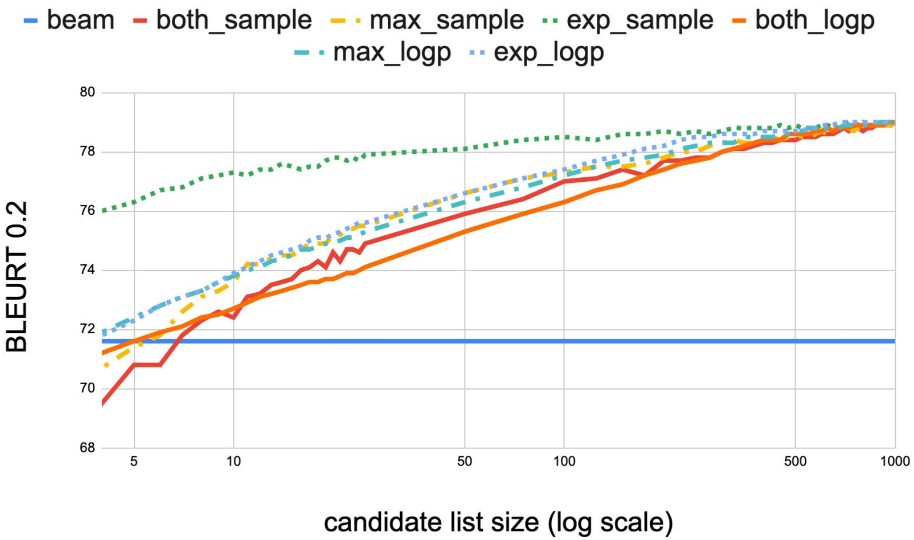
- MBR with sBLEU, chrf, or YiSi generate the same translation as beam search decoding
- MBR-BLEURT generates a better, more natural translation

No Beam Search Curse



- Quality improves when scaling the number of candidates
- Overcoming typical problems with beam search decoding

Pruning



- Randomly sampling on the expectation side is promising
- Caveat: mean BLEURT scores hide a lot of information

MBR Generates Different Translations

		Beam				MBR					Human	
		FB	O-W	UEdin	Ours	BLEU	CHRF	YISI	BL.1	BL.2	Ref-C	Ref-D
Beam	Facebook		59.5	67.6	56.9	55.6	54.0	54.1	43.3	35.0	42.0	38.4
	Online-W	59.4		56.4	53.9	52.9	52.8	51.8	42.6	34.7	41.3	40.4
	UEdin	67.6	56.5		62.1	59.5	57.4	57.8	43.7	35.4	38.0	35.7
	Ours	57.0	54.0	62.2		77.0	69.8	71.9	50.6	39.8	34.3	33.9
MBR	BLEU	55.6	53.0	59.6	77.0		73.5	76.8	50.7	40.0	34.7	33.9
	CHRF	53.9	52.8	57.4	69.7	73.4		72.1	50.6	40.0	34.2	33.1
	YISI	54.2	51.9	57.9	71.8	76.7	72.2		50.4	39.5	34.2	33.7
	BL.1	43.3	42.6	43.7	50.5	50.6	50.6	50.3		50.7	29.2	28.7
	BL.2	35.0	34.7	35.3	39.8	39.9	40.0	39.5	50.7		25.4	24.6
Human	Ref-C	42.0	41.4	38.0	34.3	34.6	34.3	34.1	29.2	25.5		31.4
	Ref-D	38.5	40.4	35.7	33.9	33.9	33.2	33.7	28.7	24.6	31.5	

- MBR-BLEURT translations are quite **<u>different</u>** compared to beam search and MBR with overlap metrics
- Similarly different like 2 different human evaluation

Conclusions

1. MBR decoding with a neural metric like BLEURT significantly outperforms beam search decoding
2. MBR decoding with BLEU outperforms beam search decoding
3. MBR-BLEURT overcomes many problems of beam search decoding:
   - Omissions errors
   - Mistranslation
   - Awkward style
   - “Beam search curse”
4. MBR-BLEURT generates translations with much lower model probabilities
   - More similar to the style of human translations
5. Many more experiments in our paper:
   - More language pairs
   - Impact of different NMT models
   - Pruning strategies
   - ...

Findings & Recommendations

Findings & Recommendations for High-quality Translations

1. **Crowd-worker human evaluation has low correlation with MQM**
    a. Unable to distinguish MT from human translation
    b. Difference in ranking of WMT submissions
    c. Bad ranking of online systems
    d. Same finding for WMT21 (check out Sec 8.3 of the Metric Task paper)
2. **Stop using crowd-worker for human evals**
    a. Unreliable, biased
    b. We have experiments comparing different rater pools based on the same human eval in the paper
3. **Use MQM**
    a. Reliable evaluation also for closer systems
    b. Error Annotation will help to understand the difference between 2 systems
    c. Error annotations should guide MT research
    d. Flexible error weighting schema
4. **Higher correlation of automatic metrics with MQM**
    a. WMT human eval correlation with MQM lower than most of the metrics
    b. MQM annotations are extremely helpful to improve and evaluate automatic metrics
        i. WMT21 and WMT22 Metric Task are using them
5. **All data is available on github:**
    a. Annotations: https://github.com/google/wmt-mqm-human-evaluation
    b. MQM viewer: https://github.com/google-research/google-research/tree/master/mqm_viewer

Future Research Directions

Human Evaluation:

1. **Establish MQM as a standard for human evaluation**
   a. Make MQM and the annotators accessible to everyone
2. **Improve Inter-annotator agreement** of raters so that we can compare MQM evaluations from different rater pools
3. **Invent new human evaluation methodologies that have high correlation with MQM, but cheaper**
   a. Is it possible to do this with crowd workers?
4. **Use MQM in other NLP fields**
   a. We expect similar findings

Automatic Evaluation:

5. **Develop trained metrics that can predict error spans and error categories**
   a. Help understand errors from systems
   b. Make MQM more accessible to everyone

MT modelling research:

6. **Develop new interesting approaches**
   a. E.g. Paragraph-level translations
7. **Re-visit existing approaches that were under-evaluated before?**
   a. Pre-training/ LM-augmented NMT

References

For papers discussed in this talk

1. APE at scale and its implications on MT evaluation biases - M Freitag, I Caswell, S Roy [WMT 19]
2. Translationese as a Language in" Multilingual" NMT - P Riley, I Caswell, M Freitag, D Grangier [ACL 19]
3. BLEU might be guilty but references are not innocent - M Freitag, D Grangier, I Caswell [EMNLP 20]
4. Results of the WMT20 metrics shared task - N Mathur, J Wei, M Freitag, Q Ma, O Bojar [WMT 20]
5. Experts, errors, and context: A large-scale study of human evaluation for machine translation - M Freitag, G Foster, D Grangier, V Ratnakar, Q Tan, W Macherey [TACL 21]
6. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain - M Freitag, R Rei, N Mathur, C Lo, C Stewart, G Foster, A Lavie, O Bojar [WMT 21]
7. Minimum Bayes Risk Decoding with Neural Metrics of Translation Quality - M Freitag, D Grangier, Q Tan, B Liang [TACL 22]
8. A Natural Diet: Towards Improving Naturalness of Machine Translation Output - M Freitag, D Vilar, D Grangier, C Cherry, G Foster [ACL 22]


شكراً جزيلاً
ngiyabonga
рахмат
danke
謝謝
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
spas
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
blagodaram
dank je
misaotra
paldies
grazzi
matondo
спасибі
kiitos
dankie
nanni
nandri
dhanyavad
maururu
thank you
gracias
хвала
asante
manana
obrigada
tenki
murakoze
köszönöm
dankon
ačiū
akun
bedankt
hvala
gràçie
bayarlalaa
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
dėkuji
sukriya
kop khun krap
obrigado
sagolun
mèsi
didi madloba
ありがとう
najis tuke
terima kasih
tanemirt
grazie
arigatô
takk
dakujem
trugarez
rahmet
kam sah hamnida
rahmat
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci