Microsoft

MT Evaluation, Human Parity

Christian Federmann

MT Marathon, Prague
September 7, 2018

Translator

MT Evaluation

MT Evaluation

Quality

- Measuring MT output quality → building quality MT output

- System building requires fast, automated metrics
  - BLEU, Meteor, TER, HTER, HyTER, BEER, …

- System analysis requires human perspective
  - Domains, scenarios, tasks

- Human parity
  - Have we lost our minds?

Automated Metrics

Automated Metrics

Capture “average quality”

- Required for system training (speed, cost)
- Corpus vs segment level accuracy

Annotator agreement

- Multiple references best

Challenges

- Reference bias
- Quality control essential

Automated Metrics

Sacre BLEU!

- Read Matt Post's paper "A call for clarity in reporting **BLEU** scores"
- Report SacreBLEU signatures in papers

Contribute

- Tokenisation for non-WMT languages needs some love

References

- Do we really evaluate against post-edited Google output?

Human Evaluation

Human Evaluation

Captures “end user perceived quality”

- Whatever that means to you
- Different for academia vs industry

Annotator agreement

- Needs to be high, otherwise useless

Challenges

- Evaluation user interfaces
- Annotator fatigue, gut feeling and mobile use

Eval Approaches

WMT early days

- Adequacy, fluency, constituent ranking, …

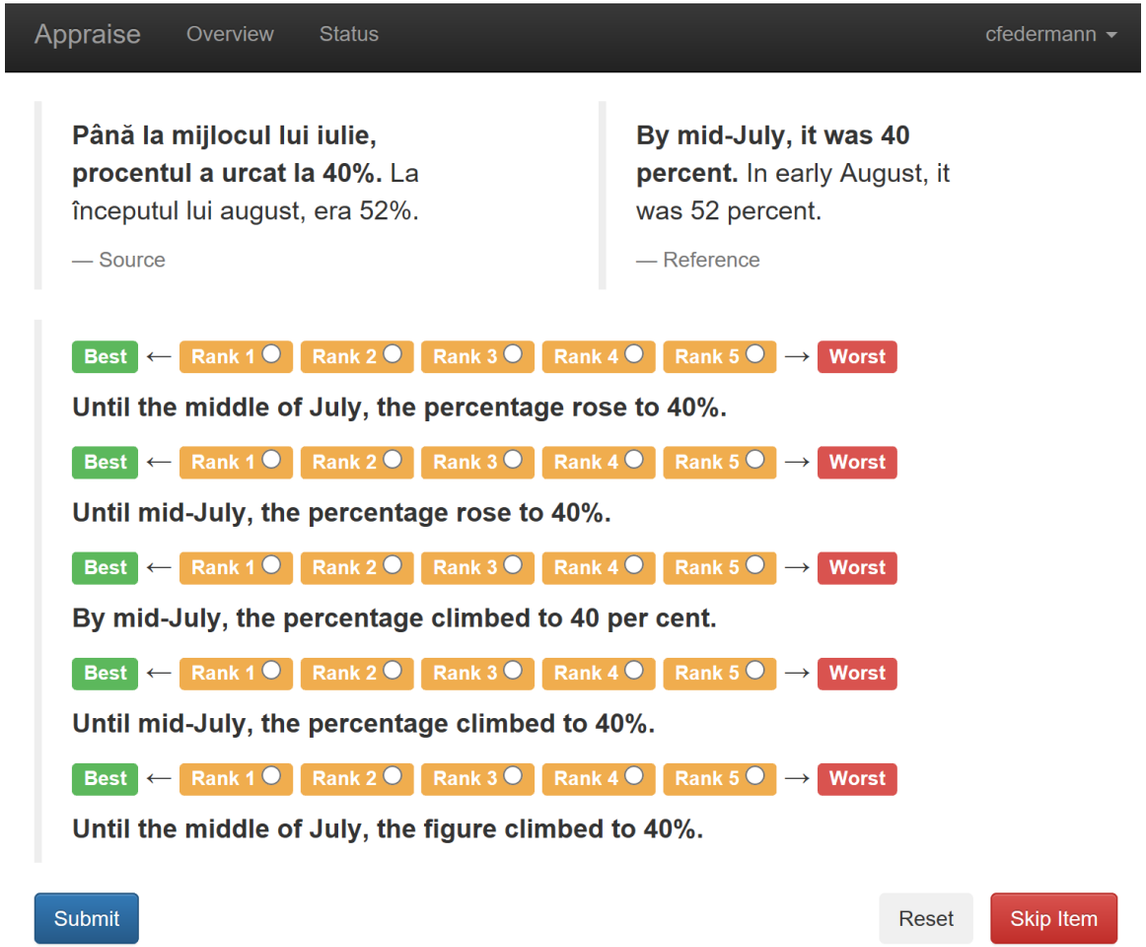
Relative ranking

- Up to five candidates per screen, ranked relative to each other
- No absolute scores, but full ranking based on TrueSkill or similar methods

Direct assessment

- Single candidate scoring, comparing to reference translation
- Adapted for source-based evaluation

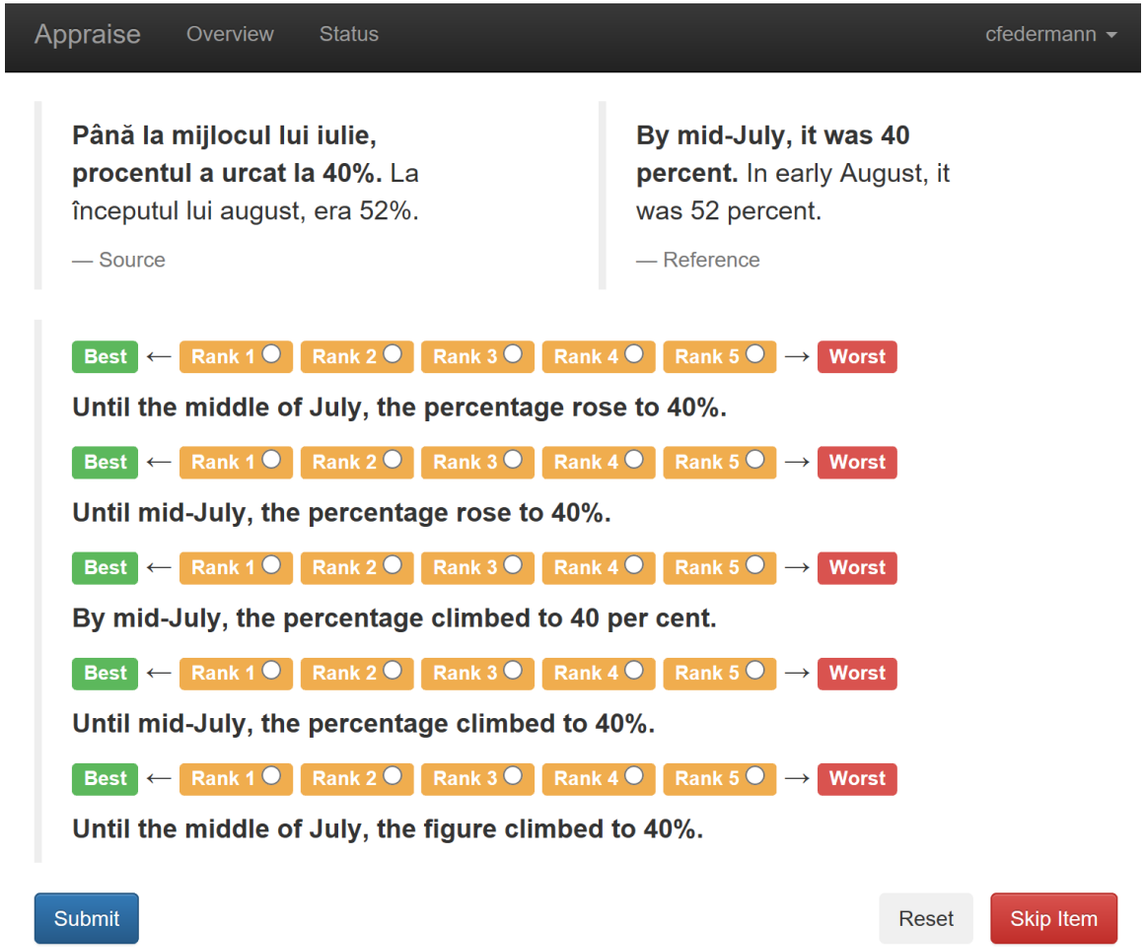
Relative Ranking



Relative Ranking

Good

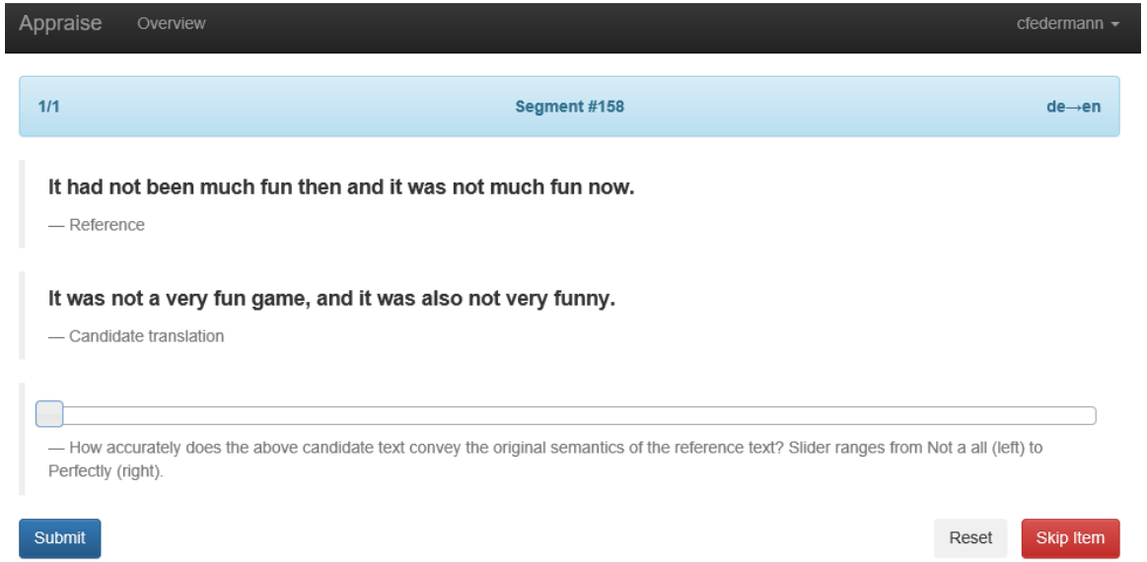
- HIT size: 3 x 5
- Relatively fast
- Skip-able
- Mental context



Bad

- Quadratic cost
- Cognitive load
- Long sentences
- Only relative deltas
- No absolute scores

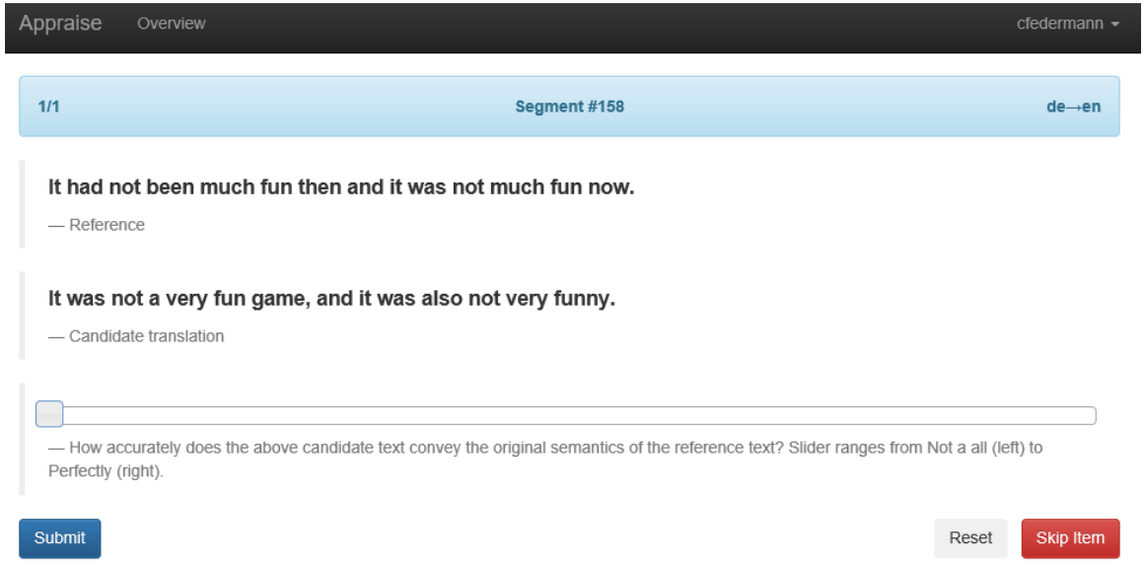
Direct Assessment



Direct Assessment

Good

- Linear cost
- Cognitive load
- Absolute scores
- Long sentences



Bad

- HIT size: 100 x 1
- Comparatively slow
- Fuzzy mental context
- High loss for crowd

Reference Bias

Shared problem

- Annotators as “human language models”
- Bad reference → bad results

Source-based?

- Annotators are “transfer raters”
- Bad source → bad results

Our solution: bilingual annotators

→ Alternative: dual references, high quality

Human Parity

"First step on the trajectory towards human parity for machine translation"

Research

Define new challenge for NMT research

- MT quality has improved a lot → how far are we from human performance?
- Fundamental question: How can we measure this?

2016 – Near Parity

- **The Verge:** *In some cases, Google says its GNMT system is even approaching human-level translation accuracy. That near-parity is restricted to transitions between related languages, like from English to Spanish and French.*

2018 – Human Parity

- Microsoft researchers achieve human parity for distant language pair Chinese to English

Defining Human Parity

Direct, equivalence-based definition

If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved human parity.

But… hard to determine "equivalence" of translation quality

Defining Human Parity

Indirect, difference-based definition

If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved human parity.

Given a reliable scoring metric, we can measure this!

Defining Human Parity

From

(Human == Machine) → Human Parity

To

¬(Human <> Machine) → Human Parity

Defining Human Parity

Assumptions

1. Possible to measure MT quality using sampled test sets
2. Possible to measure MT quality using aggregated segment scores
3. Reliable scoring metric exists

Notes

- No claim of superiority!
- Translation not necessarily error-free
- Results valid on chosen test set only

Why not use BLEU?

Automatic metrics

- Use BLEU with high quality references?
- Quality issues with original WMT references

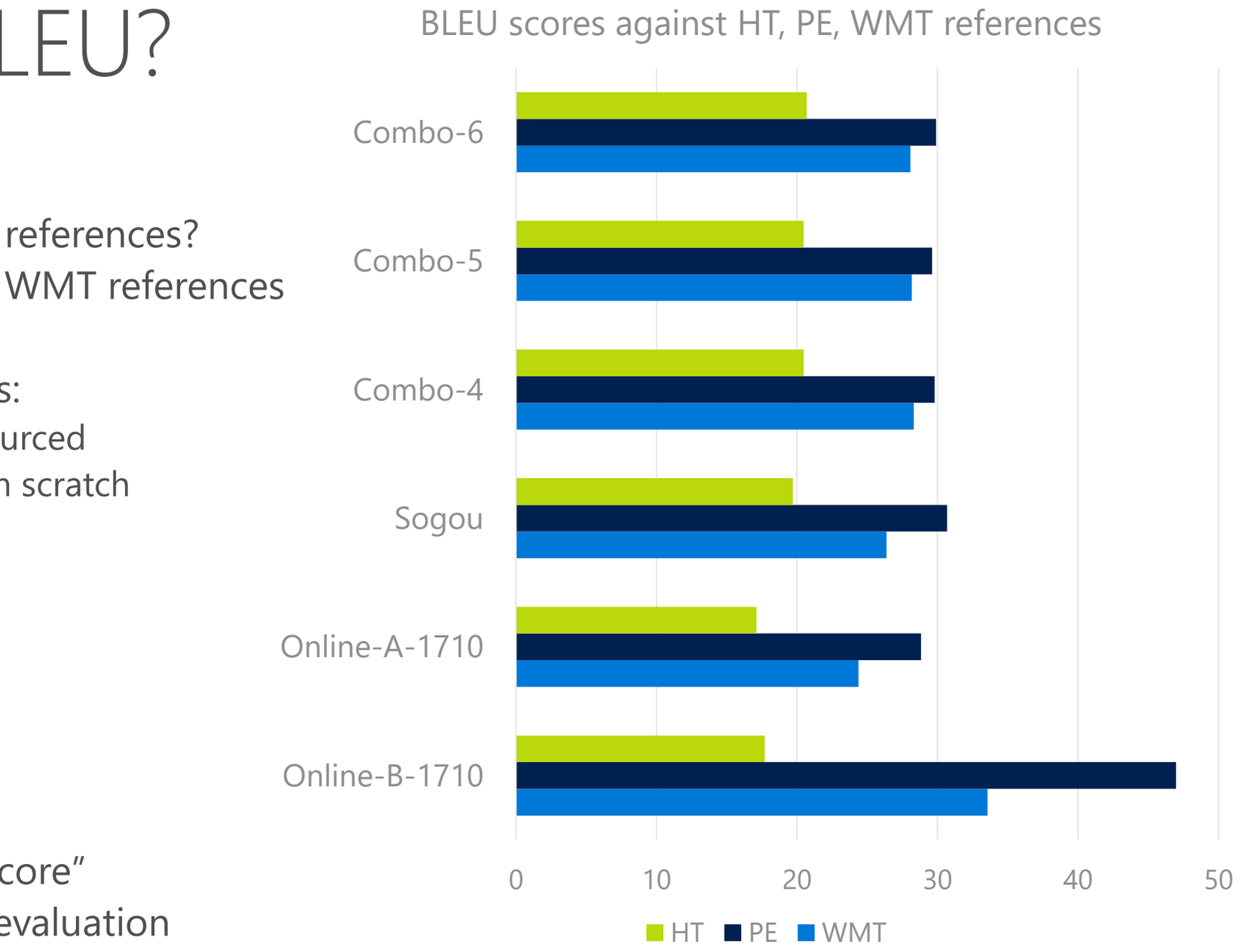
- Created two new references:
  - PE = post-edited / crowd-sourced
  - HT = human translation from scratch

Reference bias

- Online-B-1710?

Conclusions

- There is no “human BLEU score”
- Use source-based, human evaluation



Measuring Human Parity

Requirements

- Reliable scoring metric: direct assessment (DA), following state-of-the-art WMT17
- Modified to use source-based evaluation, following IWSLT17
- Enforced full overlap for all systems, with triple annotator redundancy per segment

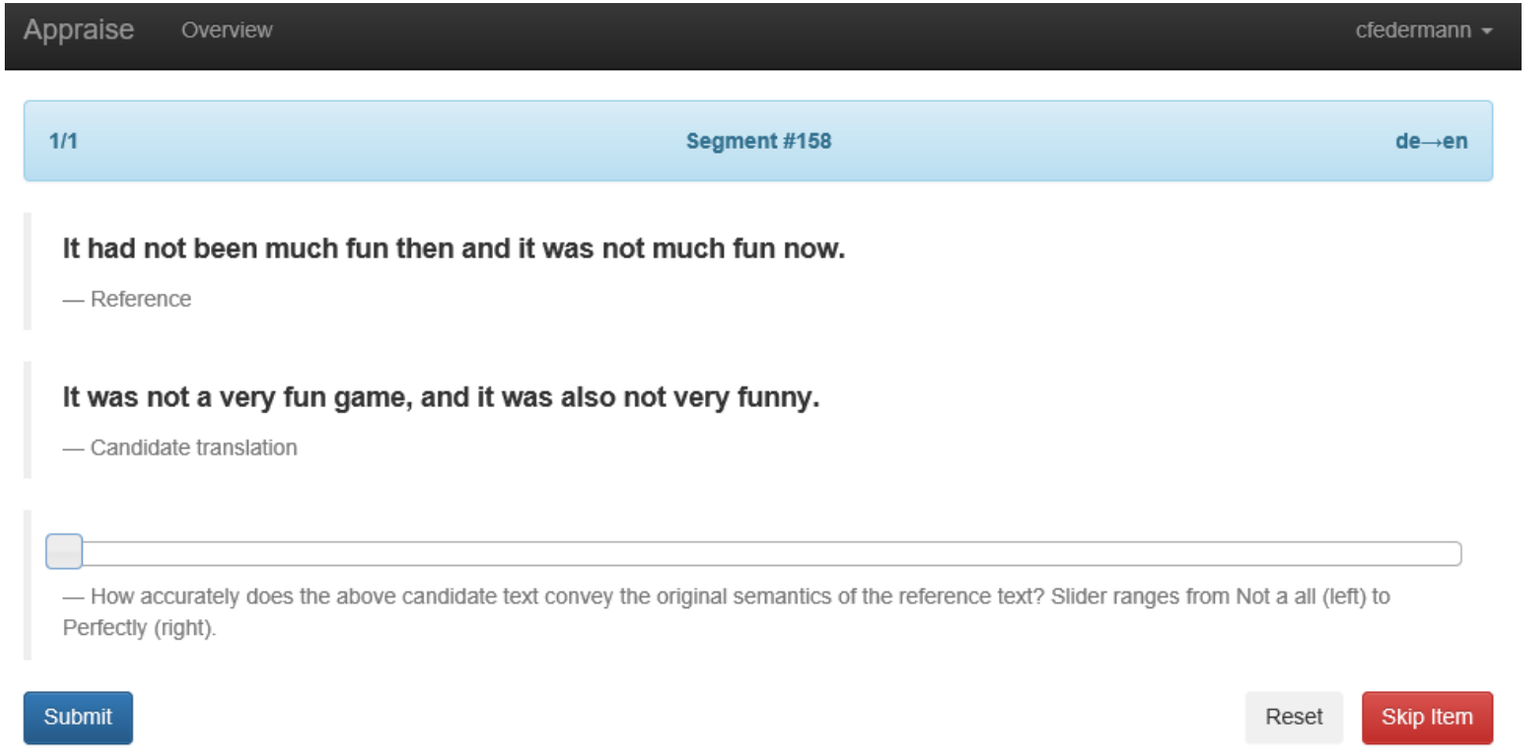
Evaluation design

- Regular evaluation campaigns over time (difference to WMT evals, which are static)
- Final evaluation campaign based on 3x Subset-1, Subset-2, Subset-3, and Subset-4
- Collected similar amount of annotations as for WMT17 → large-scale, reliable eval!
- Covering nearly half of the WMT17 test set

Direct assessment

Simple task

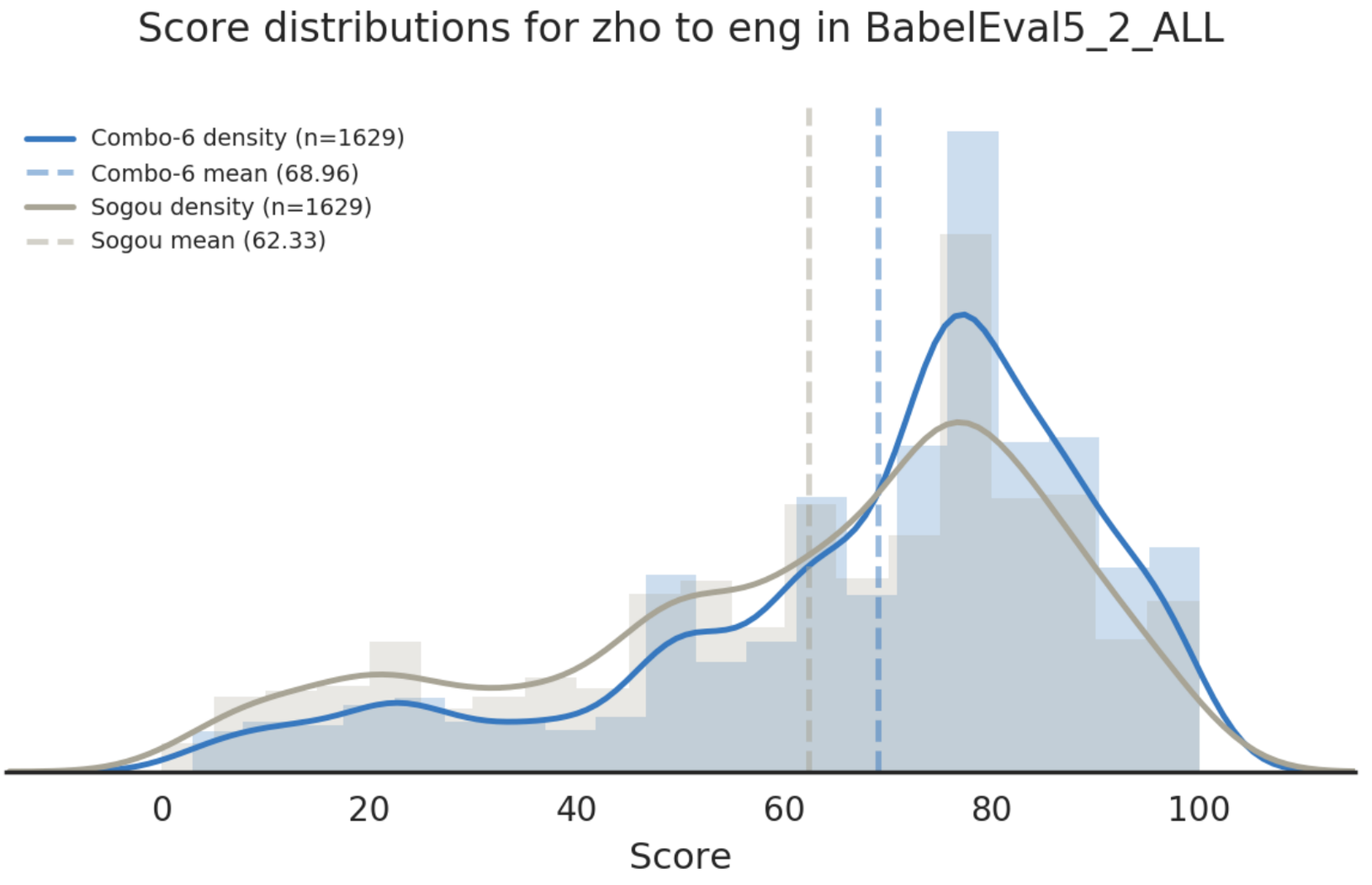
- Assigns absolute score relative to “translation hint”
- In our case, relative to source text
- Each task contains 100 items



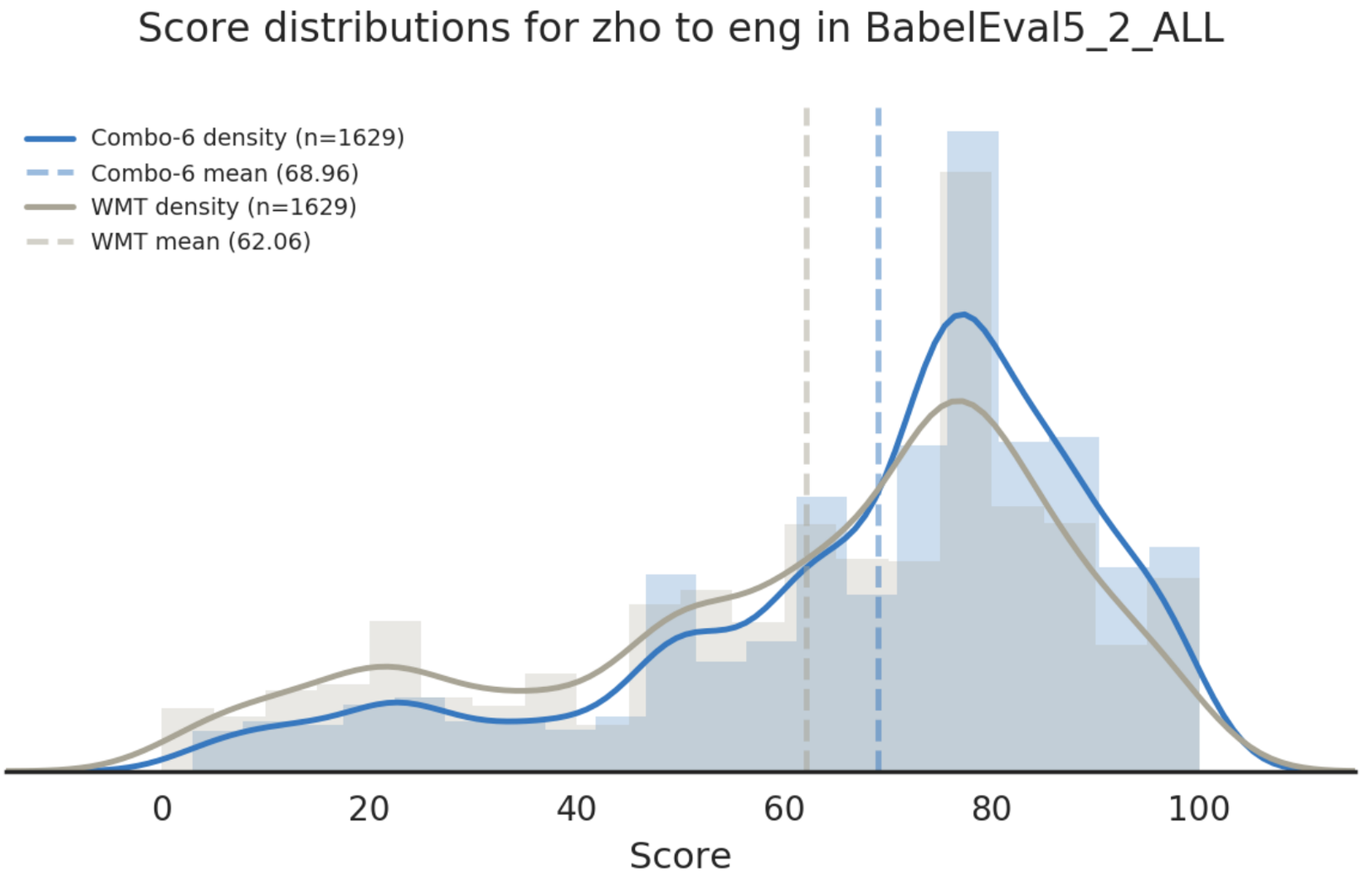
Reliable scores

- Enforced segment overlap
- Embedded quality control data
- Monitor annotator reliability

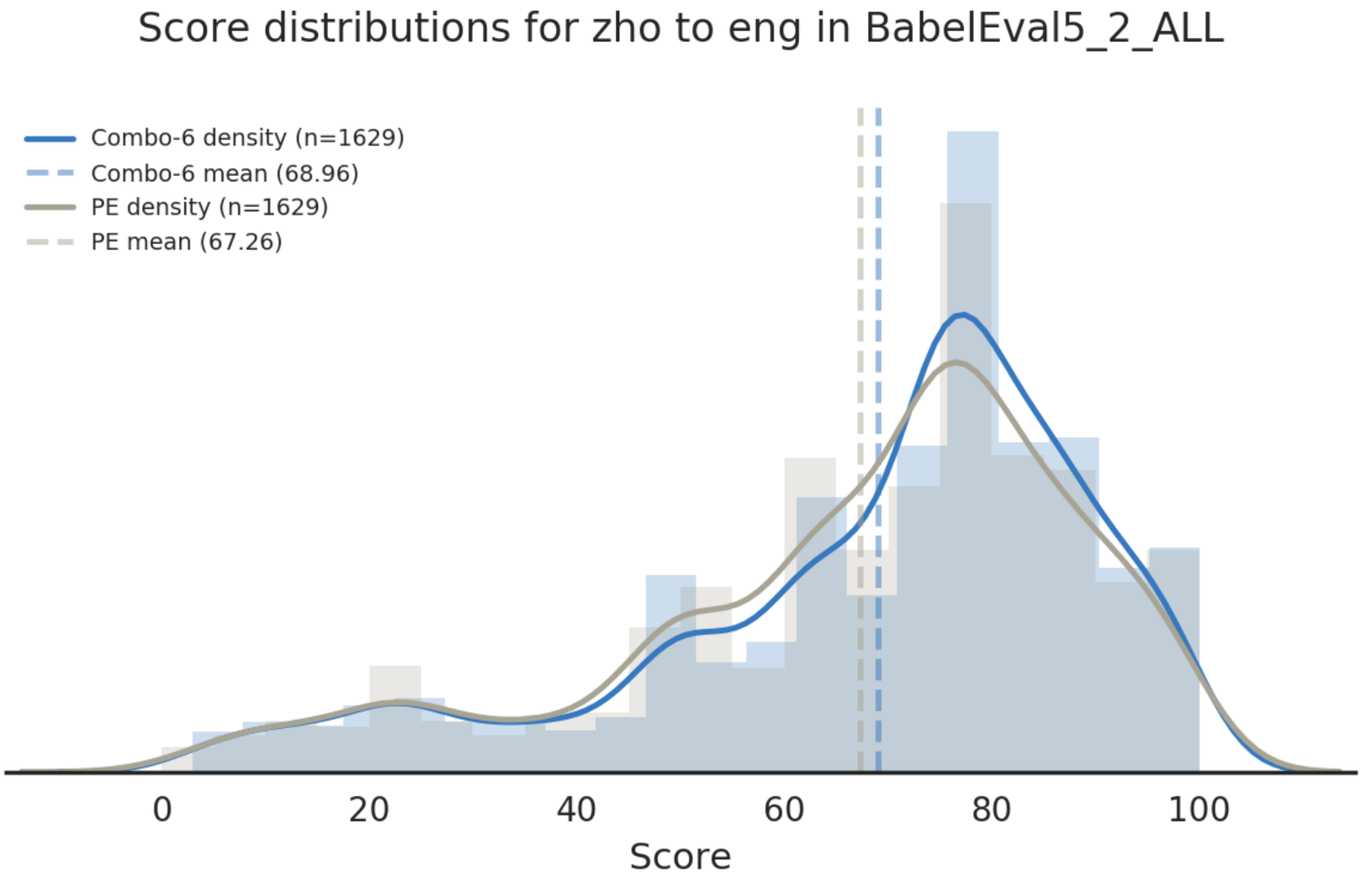
Combo-6 vs Sogou



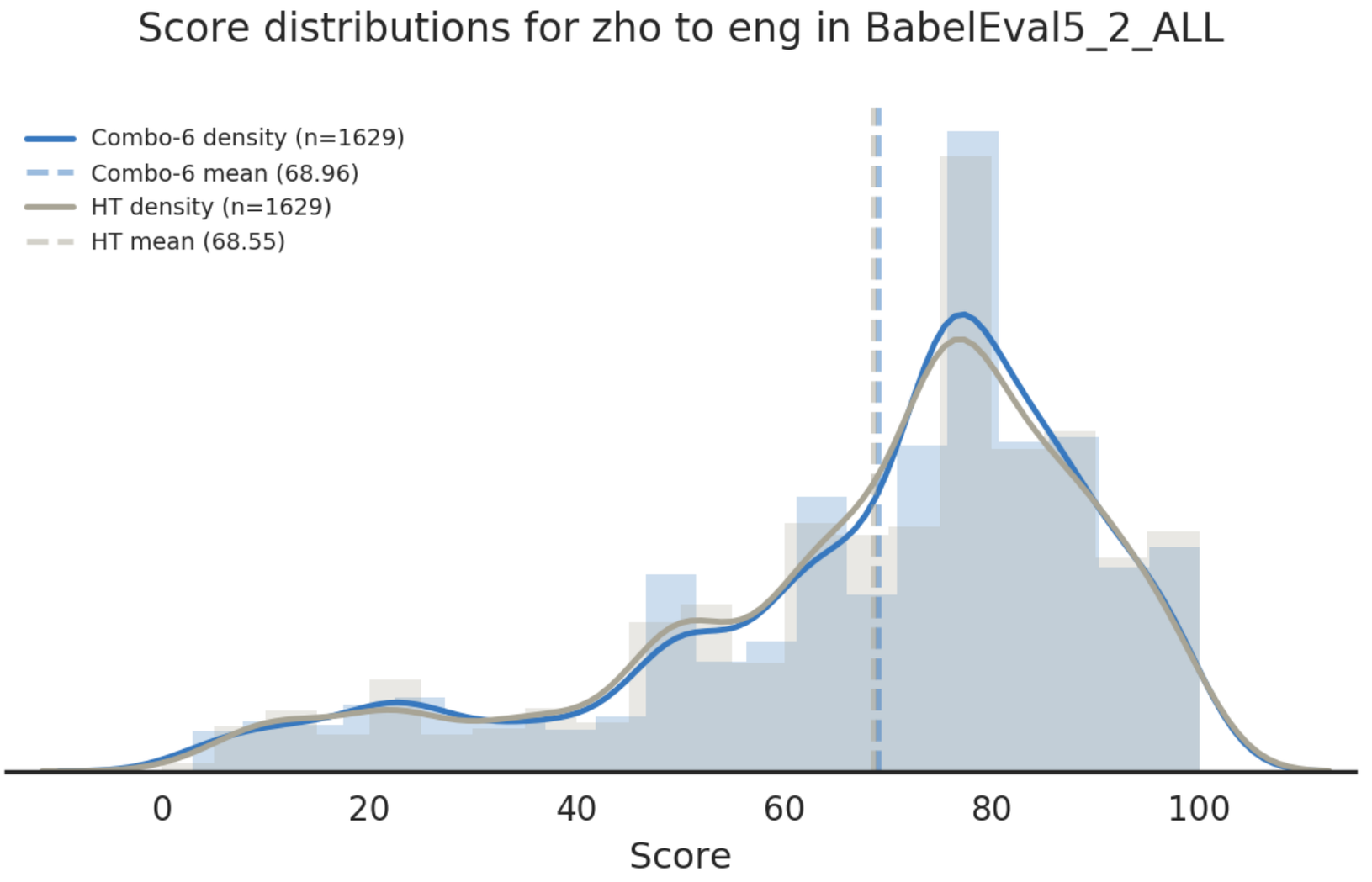
Combo-6 vs WMT



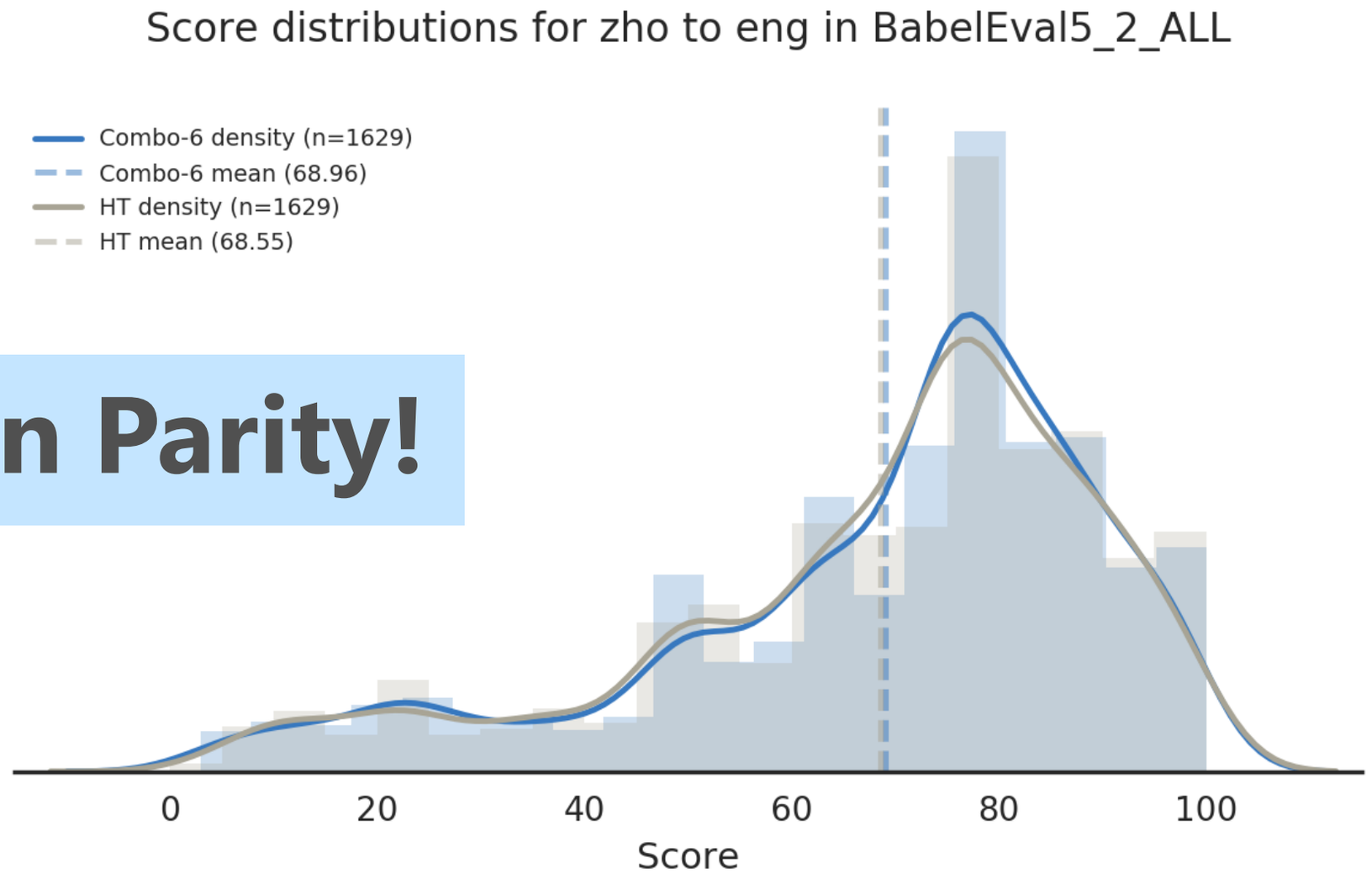
Combo-6 vs PE


Score distributions for zho to eng in BabelEval5_2_ALL
Combo-6 density (n=1629)
Combo-6 mean (68.96)
PE density (n=1629)
PE mean (67.26)
0
20
40
60
80
100
Score

Combo-6 vs HT


Score distributions for zho to eng in BabelEval5_2_ALL
Combo-6 density (n=1629)
Combo-6 mean (68.96)
HT density (n=1629)
HT mean (68.55)
0
20
40
60
80
100
Score

Combo-6 vs HT



Human Parity!

Where do we go from here?

Open data

- Released all data, including new reference translations → fostering future research
- https://github.com/MicrosoftTranslator/Translator-HumanParityData

Improved quality

- Extend human parity to consider contextual information
- Measure quality against human certification levels

Challenging future

- First step on trajectory towards human parity for machine translation
- New languages, domains, architectures

Again…

Our Definition of Human Parity…

Assumptions

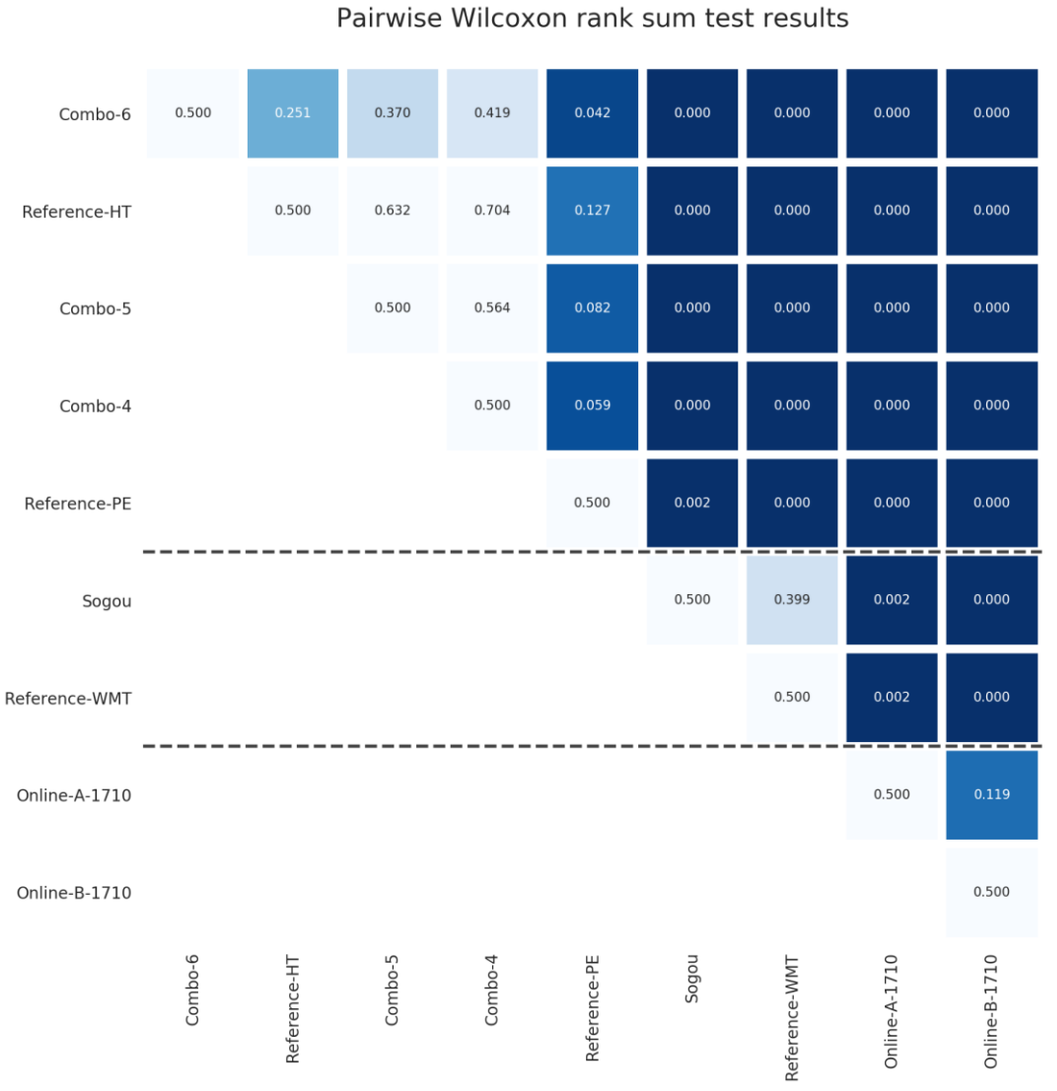
1. Possible to measure MT quality using sampled test sets
2. Possible to measure MT quality using aggregated segment scores
3. Reliable scoring metric exists

Notes

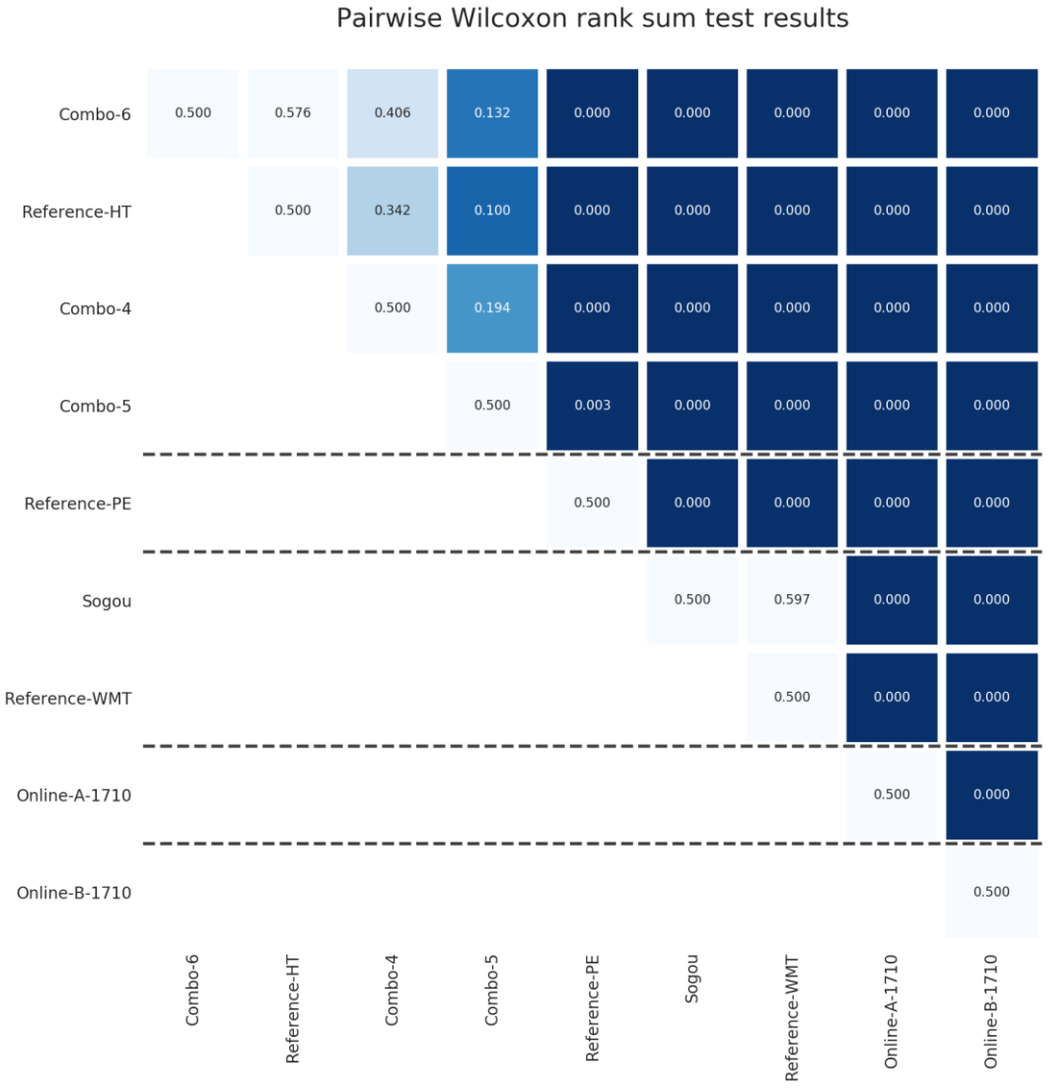
- No claim of superiority!
- Translation not necessarily error-free
- Results valid on chosen test set only

Only a start…

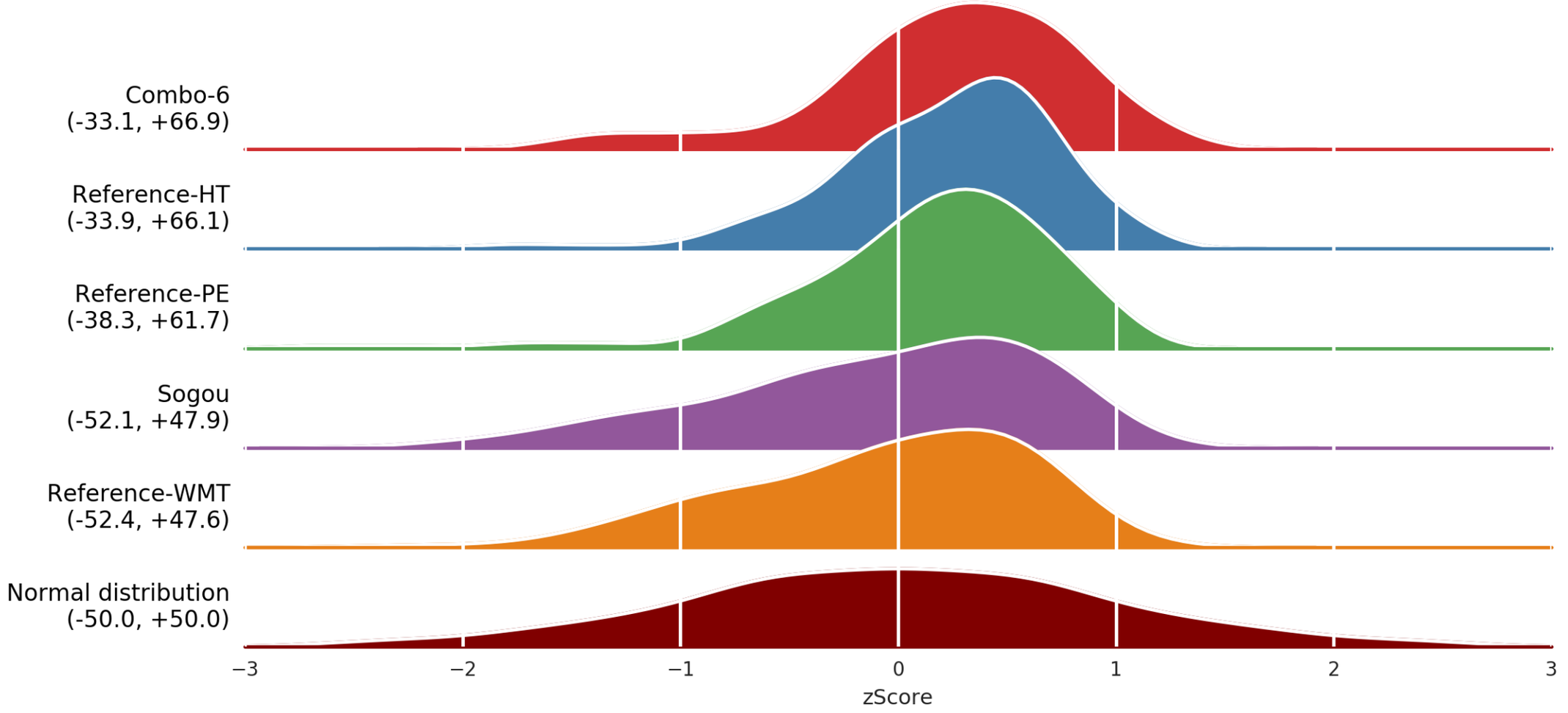
Aggregated


Pairwise Wilcoxon rank sum test results
Combo-6
Reference-HT
Combo-5
Combo-4
Reference-PE
Sogou
Reference-WMT
Online-A-1710
Online-B-1710
0.500 0.251 0.370 0.419 0.042 0.000 0.000 0.000 0.000
0.500 0.632 0.704 0.127 0.000 0.000 0.000 0.000
0.500 0.564 0.082 0.000 0.000 0.000 0.000
0.500 0.059 0.000 0.000 0.000 0.000
0.500 0.002 0.000 0.000 0.000
0.500 0.399 0.002 0.000
0.500 0.002 0.000
0.500 0.119
0.500

Individual


Pairwise Wilcoxon rank sum test results
Combo-6
Reference-HT
Combo-4
Combo-5
Reference-PE
Sogou
Reference-WMT
Online-A-1710
Online-B-1710
0.500 0.576 0.406 0.132 0.000 0.000 0.000 0.000 0.000
0.500 0.342 0.100 0.000 0.000 0.000 0.000 0.000
0.500 0.194 0.000 0.000 0.000 0.000 0.000
0.500 0.003 0.000 0.000 0.000 0.000
0.500 0.000 0.000 0.000 0.000
0.500 0.597 0.000 0.000
0.500 0.000 0.000
0.500 0.000
0.500

What is Human Parity?



Dissecting Babel Results

Criticism

Directionality

- Half of all segments are based on human translation of native English

Quality

- Reference-HT quality questionable for some segments

Context

- Evaluation only considers segment level quality → This followed the WMT17 “state of the art”

Impact

Full data

- 2,001 segments

Directionality

- 1,001 segments based on translated English; 1,000 segments based on native Chinese

Quality

- 194 segments with human DA scores ≤ 50; 1,807 segments with scores >50

Notes

- This is a rather extreme good/bad classification
- We will look into DA scores ≤ 25 for re-translation to get improved Reference-HT’

Combined impact

Scale

- 2,001 segments
- 1,001 segments based on translated English; 1,000 segments based on native Chinese
- 194 segments with human DA scores ≤ 50; 1,807 segments with scores >50

Source language	Low Ref-HT quality	High Ref-HT quality
Translated English	5%	45%
Native Chinese	5%	45%

Source: translated English vs native Chinese

EN

Rank	Z score	R score	System ID
1	0.433	73.7	Combo-6
	0.400	72.9	Combo-4
	0.391	73.1	Combo-5
	0.265	70.5	Reference-PE
	0.253	70.0	Reference-HT
	0.166	68.7	Sogou
2	-0.088	63.0	Reference-WMT
	-0.211	60.0	Online-B-1710
	-0.217	61.1	Online-A-1710

ZH

Rank	Z score	R score	System ID
1	0.187	67.1	Reference-HT
	0.047	64.4	Combo-6
	0.047	64.8	Combo-5
	0.030	64.4	Combo-4
	0.022	64.1	Reference-PE
	-0.142	61.1	Reference-WMT
2	-0.346	56.1	Sogou
3	-0.573	51.1	Online-A-1710
	-0.716	48.4	Online-B-1710

Reference: low vs high quality

LQ

Rank	Z score	R score	System ID
1	0.102	68.6	Combo-6
	0.061	68.4	Combo-5
	0.052	65.7	Reference-PE
	0.016	65.6	Reference-WMT
	-0.083	64.8	Combo-4
	-0.115	64.3	Reference-HT
	-0.447	58.3	Sogou
	-0.640	54.1	Online-A-1710
	-0.680	53.5	Online-B-1710

HQ

Rank	Z score	R score	System ID
1	0.256	69.2	Combo-6
	0.245	68.9	Reference-HT
	0.234	69.0	Combo-4
	0.233	69.1	Combo-5
	0.134	67.2	Reference-PE
2	-0.064	62.7	Sogou
	-0.127	62.0	Reference-WMT
3	-0.358	56.5	Online-A-1710
	-0.452	54.1	Online-B-1710

Combined results for translated English

EN+LQ

Rank	Z score	R score	System ID
1	0.436	74.5	Combo-5
	0.393	73.3	Combo-6
	0.340	72.3	Combo-4
	0.233	67.7	Reference-PE
	0.094	65.6	Reference-HT
	0.046	67.6	Sogou
	0.041	62.8	Reference-WMT
	0.010	63.8	Online-B-1710
	-0.442	55.3	Online-A-1710

EN+HQ

Rank	Z score	R score	System ID
1	0.453	74.4	Combo-6
	0.409	73.4	Combo-4
	0.392	73.4	Combo-5
	0.253	70.7	Reference-PE
	0.241	70.0	Reference-HT
	0.163	68.5	Sogou
2	-0.118	63.0	Reference-WMT
	-0.170	62.0	Online-A-1710
	-0.245	59.3	Online-B-1710

Combined results for native Chinese

ZH+LQ

Rank	Z score	R score	System ID
1	-0.007	68.0	Reference-WMT
	-0.106	63.9	Reference-PE
	-0.152	64.5	Combo-6
	-0.266	63.0	Combo-5
	-0.298	63.2	Reference-HT
	-0.452	58.2	Combo-4
	-0.813	53.0	Online-A-1710
	-0.879	50.2	Sogou
	-1.282	44.5	Online-B-1710

ZH+HQ

Rank	Z score	R score	System ID
1	0.250	67.8	Reference-HT
	0.080	64.7	Combo-5
	0.068	64.9	Combo-6
	0.066	64.2	Combo-4
	0.019	63.9	Reference-PE
	-0.135	61.0	Reference-WMT
2	-0.281	57.1	Sogou
3	-0.540	51.2	Online-A-1710
	-0.651	49.1	Online-B-1710

Combined results for all data

ALL

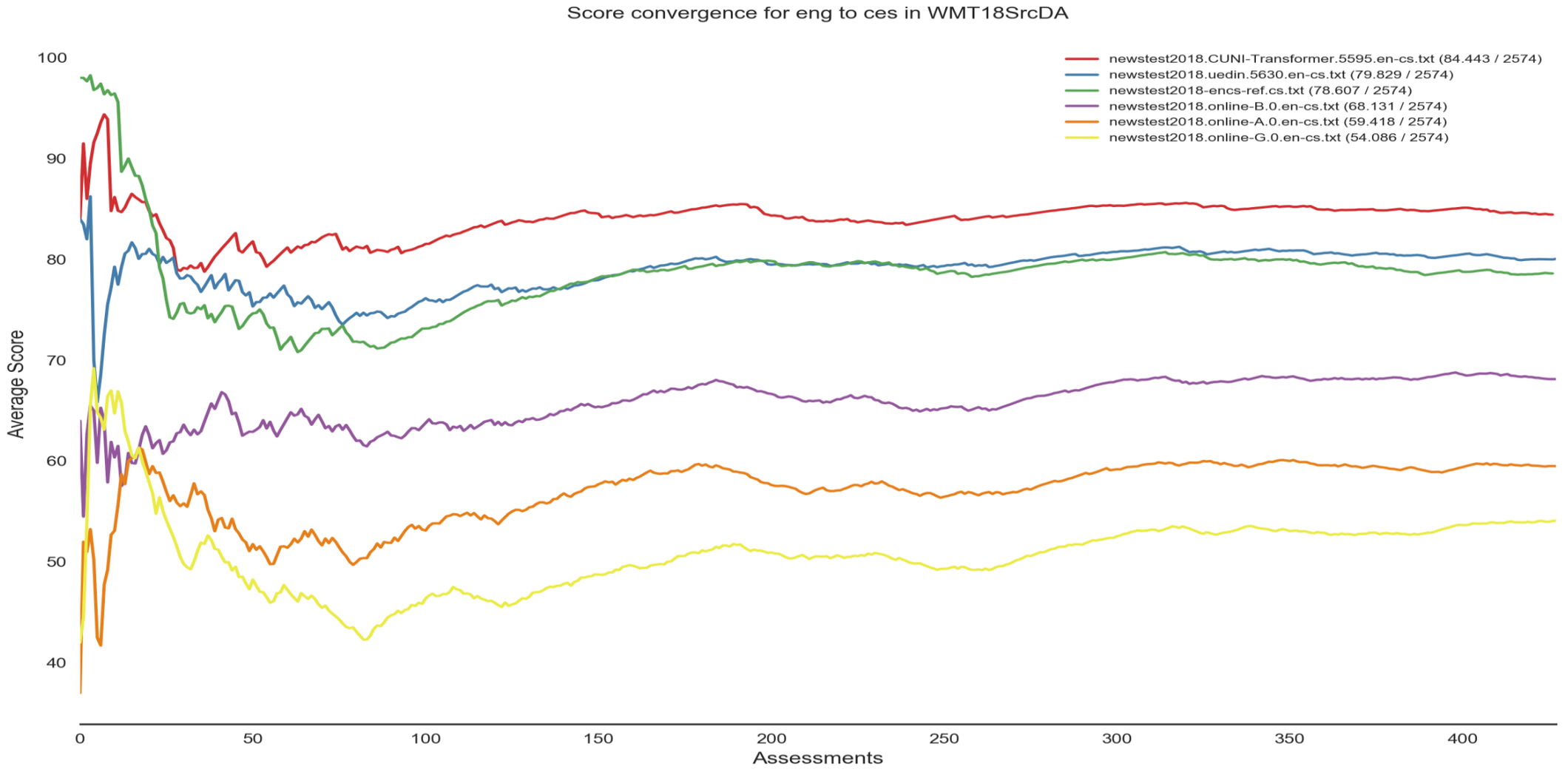
Rank	Z score	R score	System ID
1	0.237	69.0	Combo-6
	0.220	68.5	Reference-HT
	0.216	68.9	Combo-5
	0.211	68.6	Combo-4
	0.141	67.3	Reference-PE
2	-0.094	62.3	Sogou
	-0.115	62.1	Reference-WMT
3	-0.398	56.0	Online-A-1710
	-0.468	54.1	Online-B-1710

Conclusions

- Human parity for all data subsets
- Research systems do well on translated English source text
- Reference-HT should be improved for LQ data subset

Appendix

WMT18 English to Czech (convergence)



WMT18 English to Czech (scores)