Human Evaluation and Human Parity for Chinese-English News Translation

Christian Federmann
Microsoft Translator

Microsoft Translator and MSRA Teams

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, **Shujie Liu**, **Tie-Yan Liu**, Renqian Luo, Arul Menezes, **Tao Qin**, Frank Seide, **Xu Tan**, **Fei Tian**, Lijun Wu, Shuangzhi Wu, Yingce Xia, **Dongdong Zhang**, Zhirui Zhang, Ming Zhou

arxiv paper: https://arxiv.org/abs/1803.05567

“First step on the trajectory towards human parity for machine translation”

Project Babel: a roadmap to Human Parity

- 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

Outrageous!

Focus on

Human evaluation

Goal

Measure human parity

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 reference

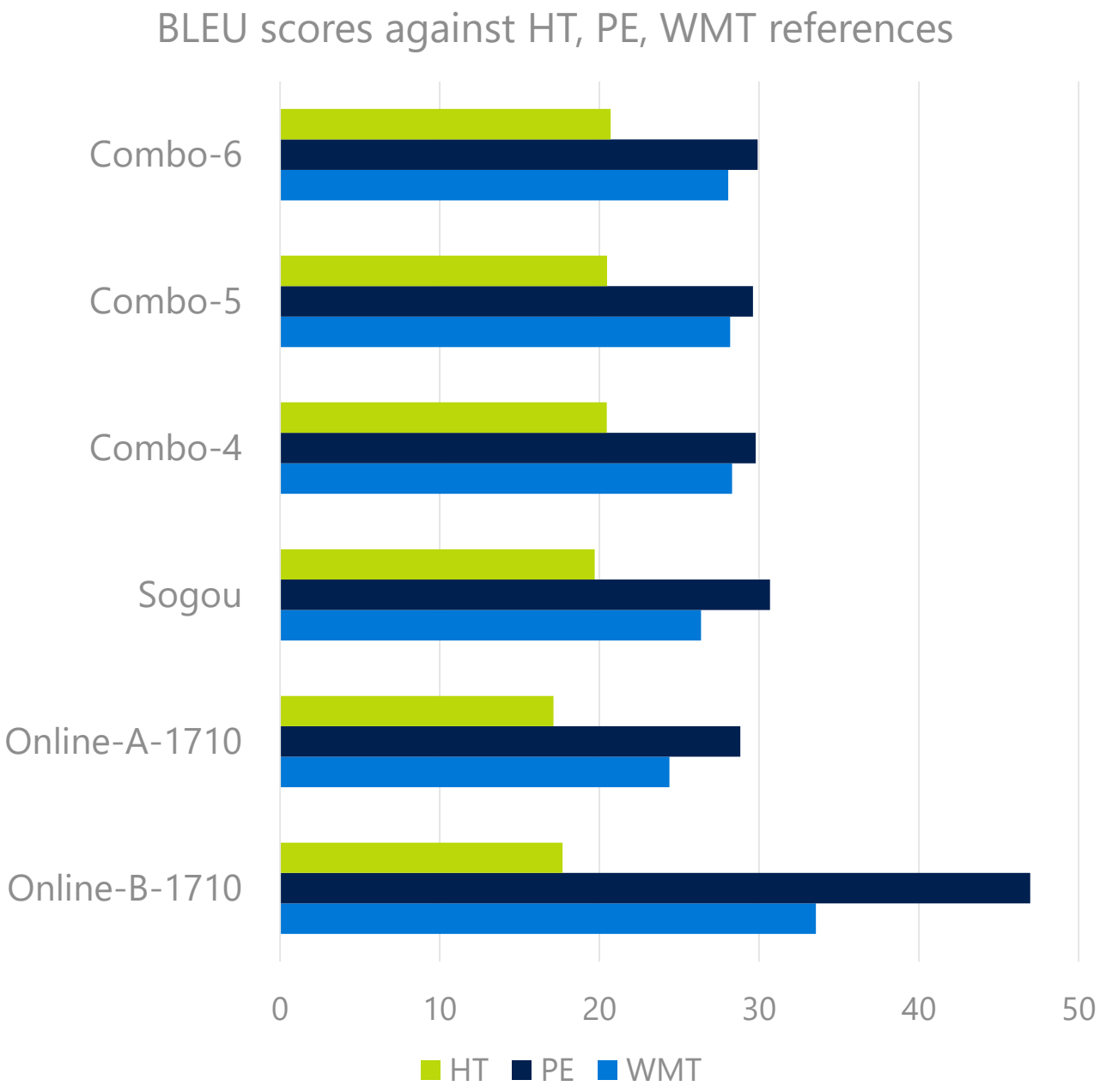
- Created two new references:
  - PE = post-edited / crowd-sourced
  - HT = human translation from scratch

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

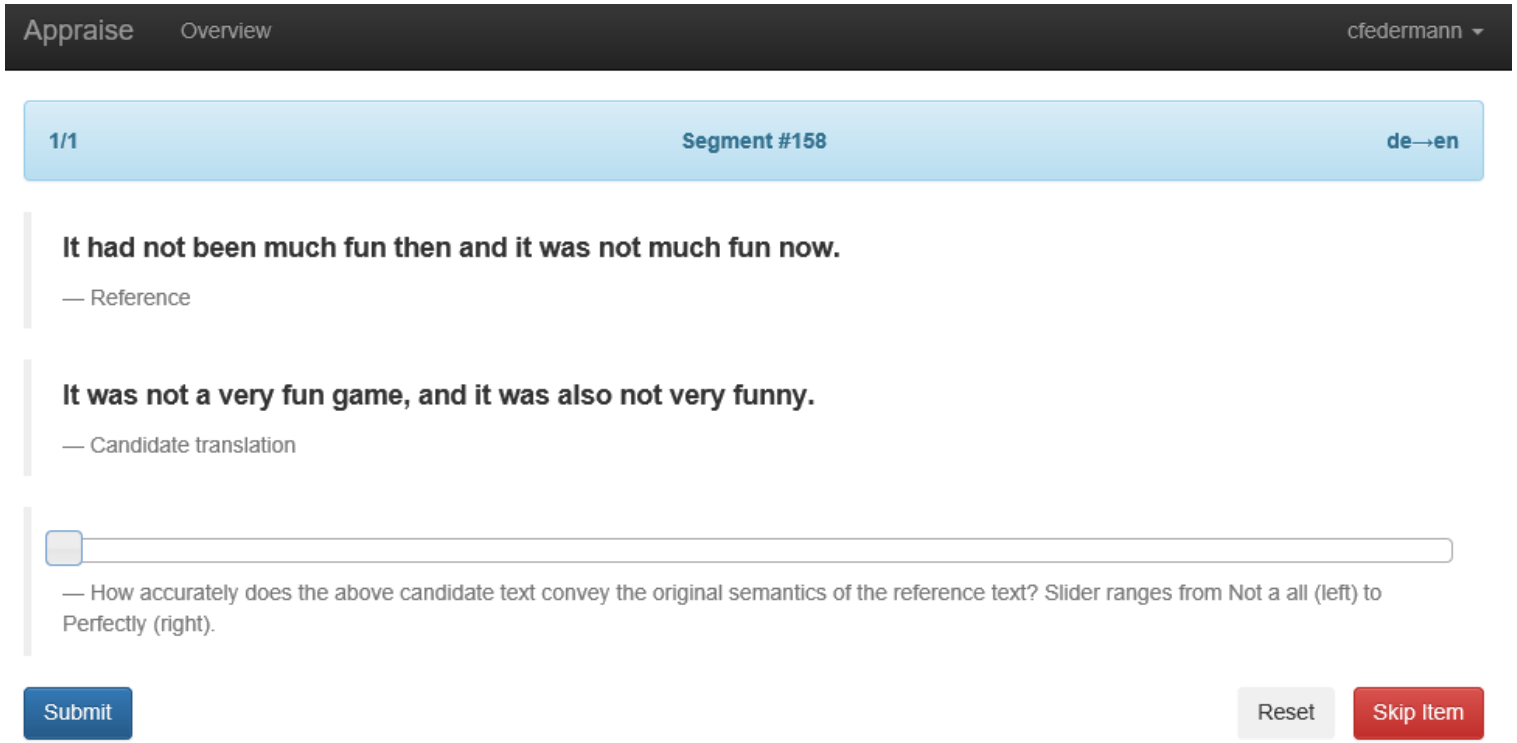
Direct assessment

Simple task

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

Reliable scores

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



Campaigns

Timeline

- Regular monthly evaluation campaigns
- Final round of evaluation campaigns in February/March 2018

Scale

- 9 systems under investigation, including 3 research candidates
- 15 annotators per subset, 6 subsets
- 20 tasks per subset, 3 redundant annotators per task
- 4,200 data points per subset (excluding quality controls)
- 25,200 data points across all subsets

Project Babel: Monthly evaluations

November 2017	Rank group
Pactera human translation	1
Unbabel post-edited	1
Sogou Knowing NMT	2
MSR 20171012 research	**3**
Online-A-1710	4
Online-B-1710	4

December 2017	Rank group
Unbabel post-edited	1
Pactera human translation	1
Dual Learning TF/Transformer	**2**
Transformer+R2L	**2**
Karnak/Transformer	**2**
Sogou Knowing NMT	3
Dual Learning Karnak/NMT	**3**
TF/Transformer	**4**
Transformer+R2L+BackTrans	**4**
Online-B-1710	5
Online-A-1710	6

January 2018	Rank group
Unbabel post-edited	1
Pactera human trans	2
MSR Redmond 20180112	**3**
MSRA NLC 20180108	**3**
Sogou Knowing NMT	4
MSRA ML 20180111	**4**
Online-A-1710	4
Online-B-1710	5

February 2018	Rank group
Pactera human trans	1
Unbabel post-edited	1
MSRA ML 20180212	**2**
MSR Redmond 20180212	**2**
MSRA NLC 20180211	**3**
Sogou Knowing NMT	4
newstest 2017 reference	4
Online-A-1710	5
Online-B-1710	5

Evaluation against fixed-points: Two human translations, online MT systems, and Sogou

Sogou is winner of WMT2017 competition

Evaluating various research systems (***italics***)

Result clusters

Tabular representation

- Clustering boils down to pairwise differences and their significance
- Clusters based on number of significant wins against all lower ranked systems
- Systems within same cluster are considered indistinguishable
- Wilcoxon rank sum test

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

Visualising Human Parity

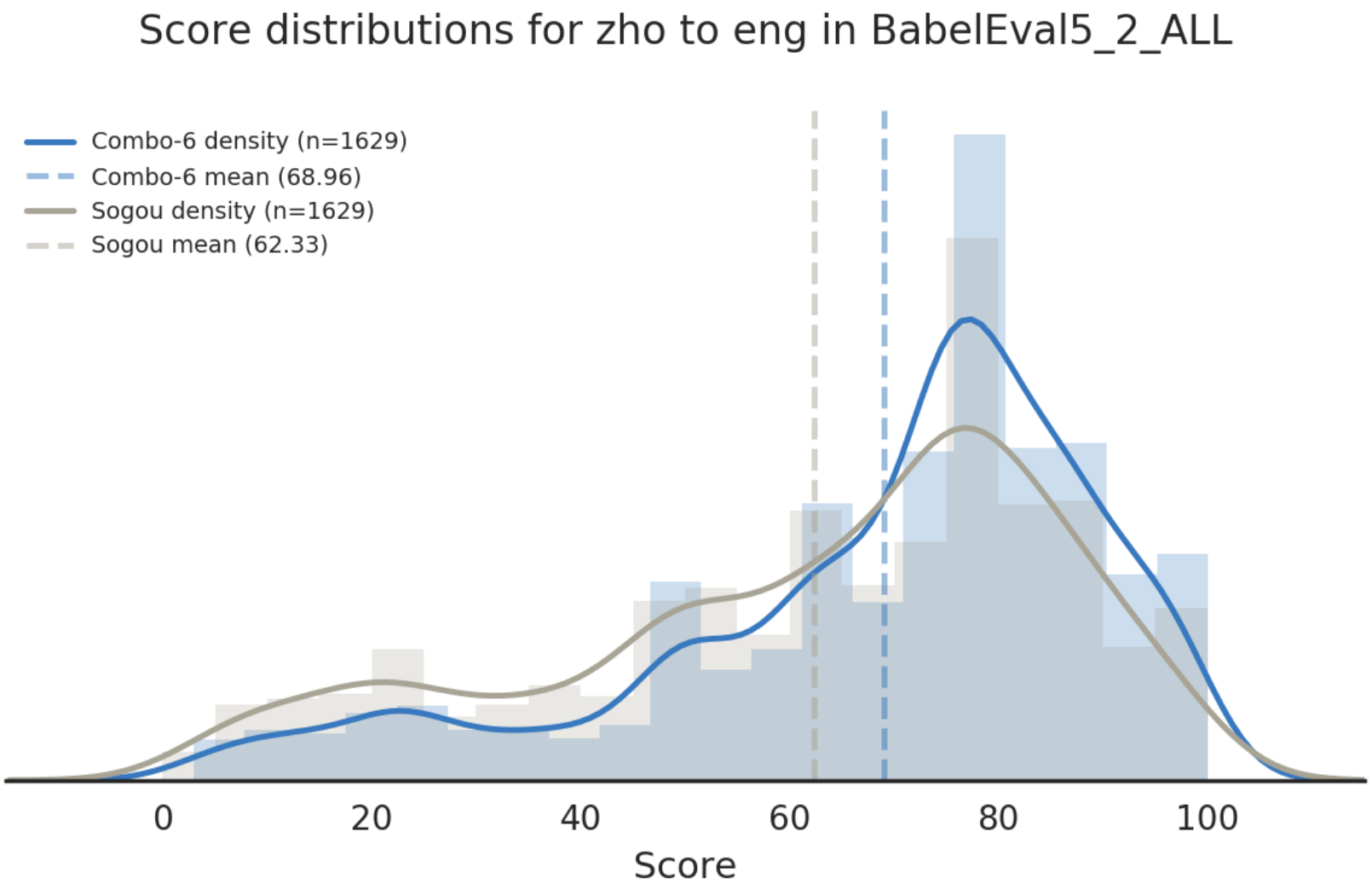
From

Result clusters for all systems

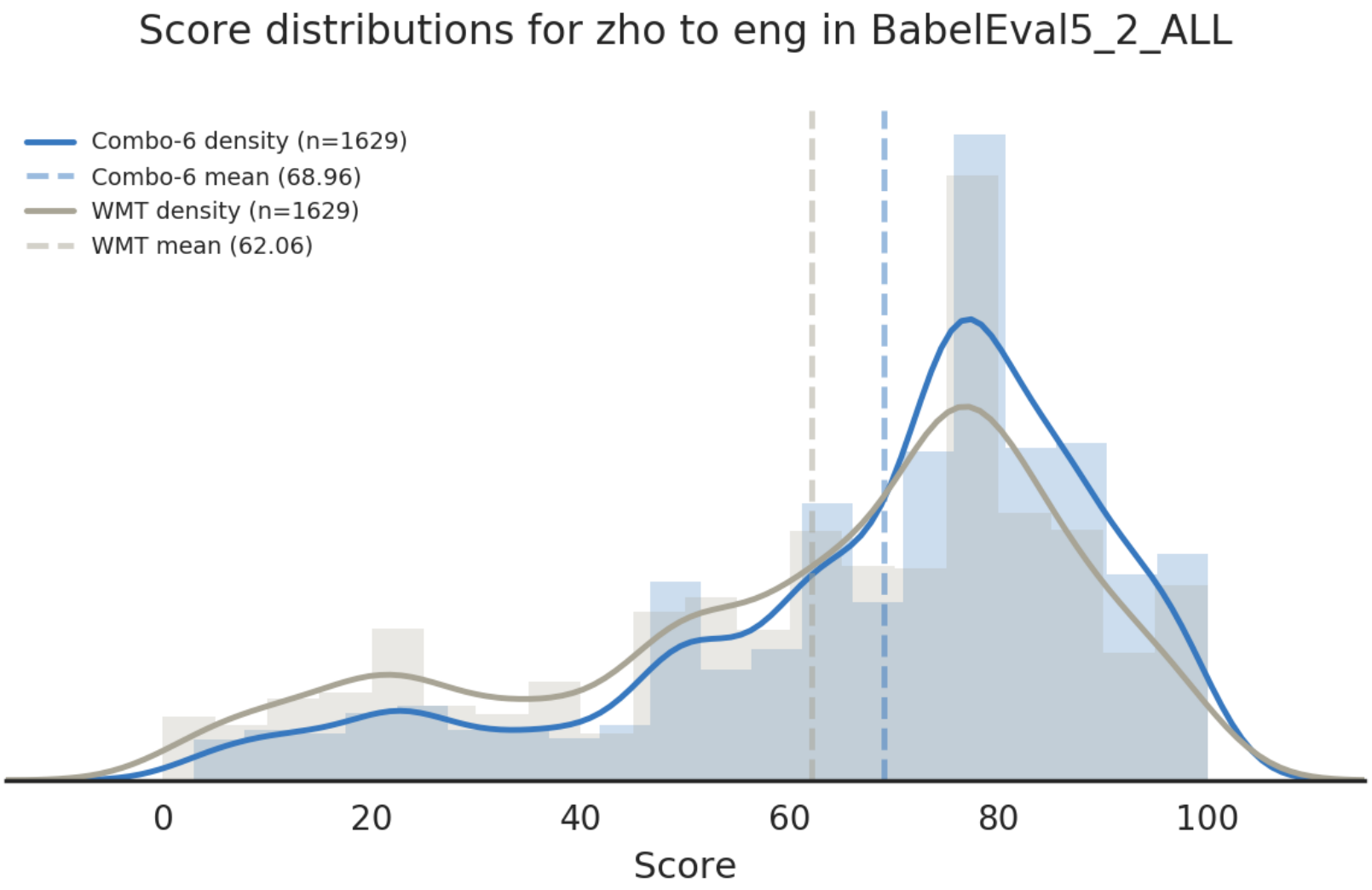
To

Pairwise density representation

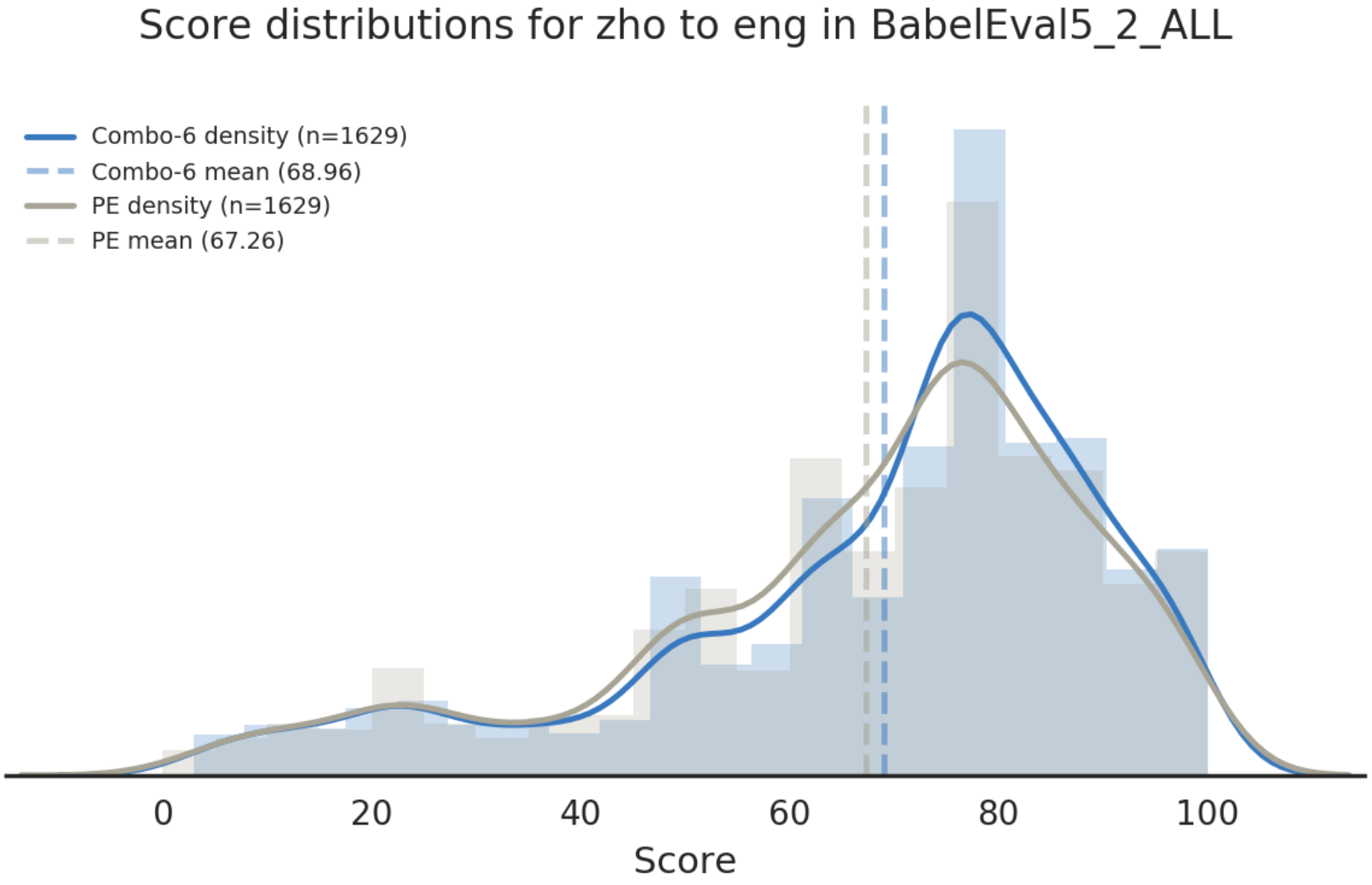
Combo-6 vs Sogou



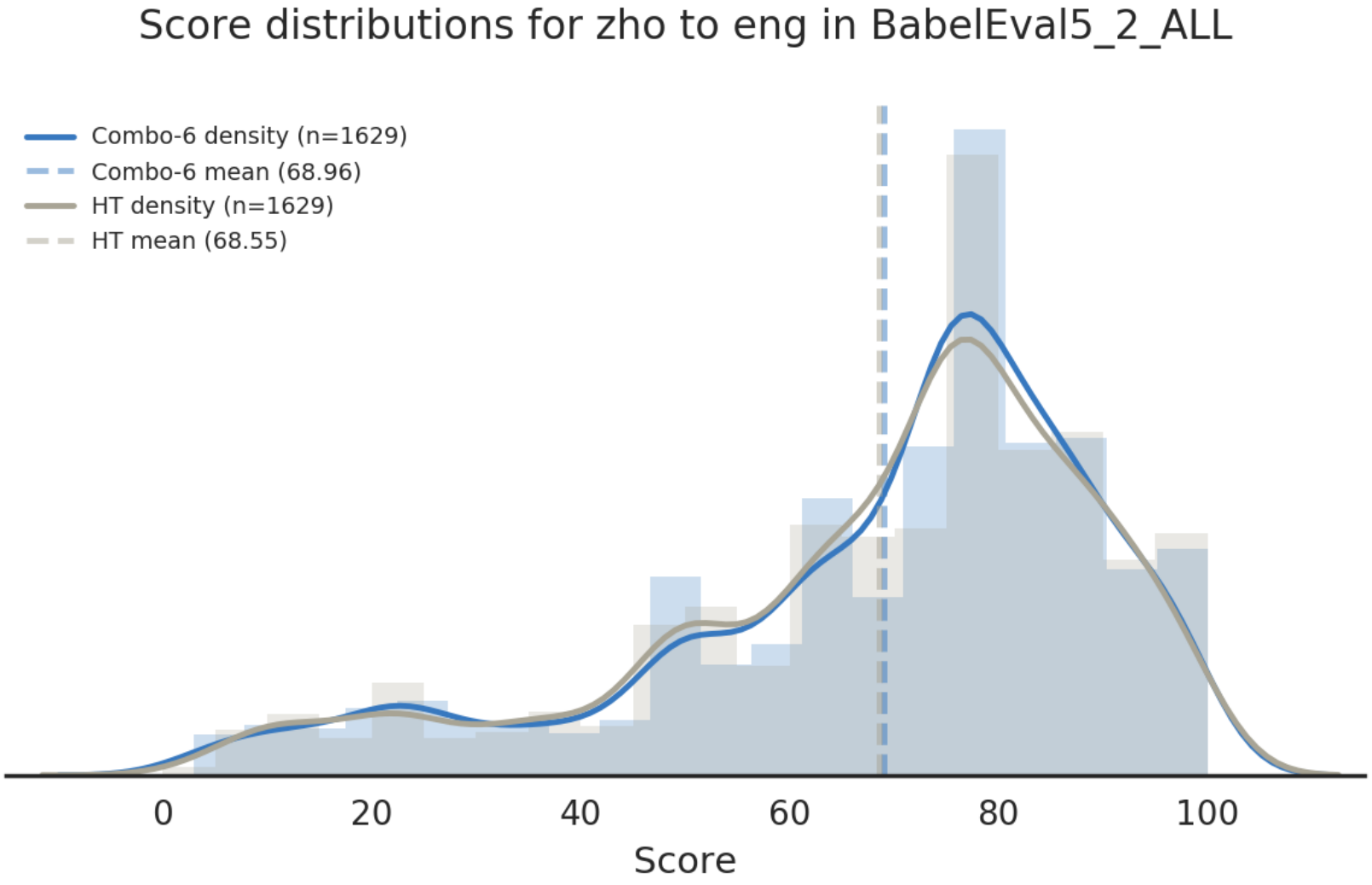
Combo-6 vs WMT



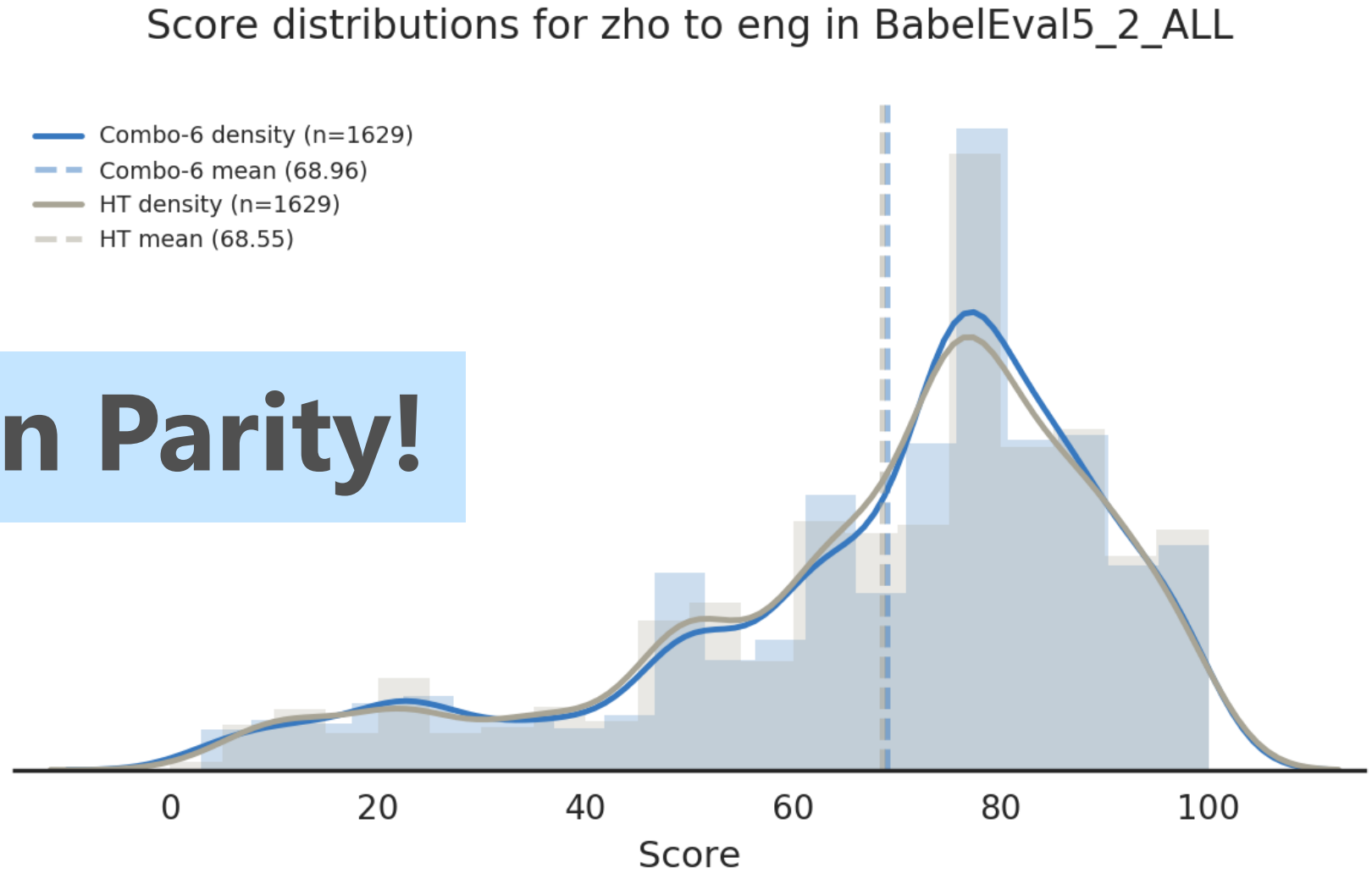
Combo-6 vs PE


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

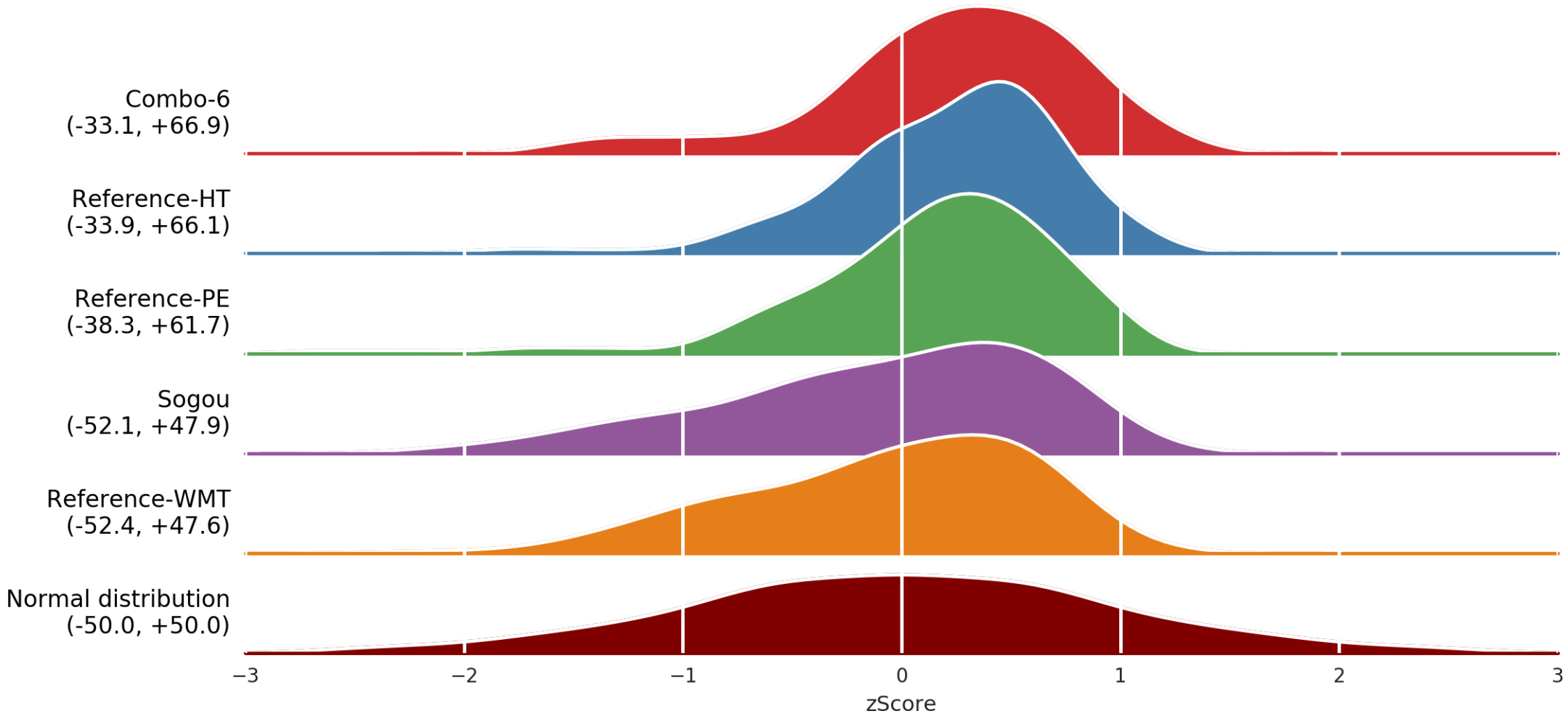

Score distributions for zho to eng in BabelEval5_2_ALL
Combo-6 density (n=1629)
Combo-6 mean (68.96)
HT density (n=1629)
HT mean (68.55)
0
20
40
60
80
100
Score

Combo-6 vs HT



Human Parity!

Combined view



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