Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

Microsoft

Found in Translation: Achieving Human Parity on Chinese-English News Translation

Hany Hassan Awadalla
Christian Federmann

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

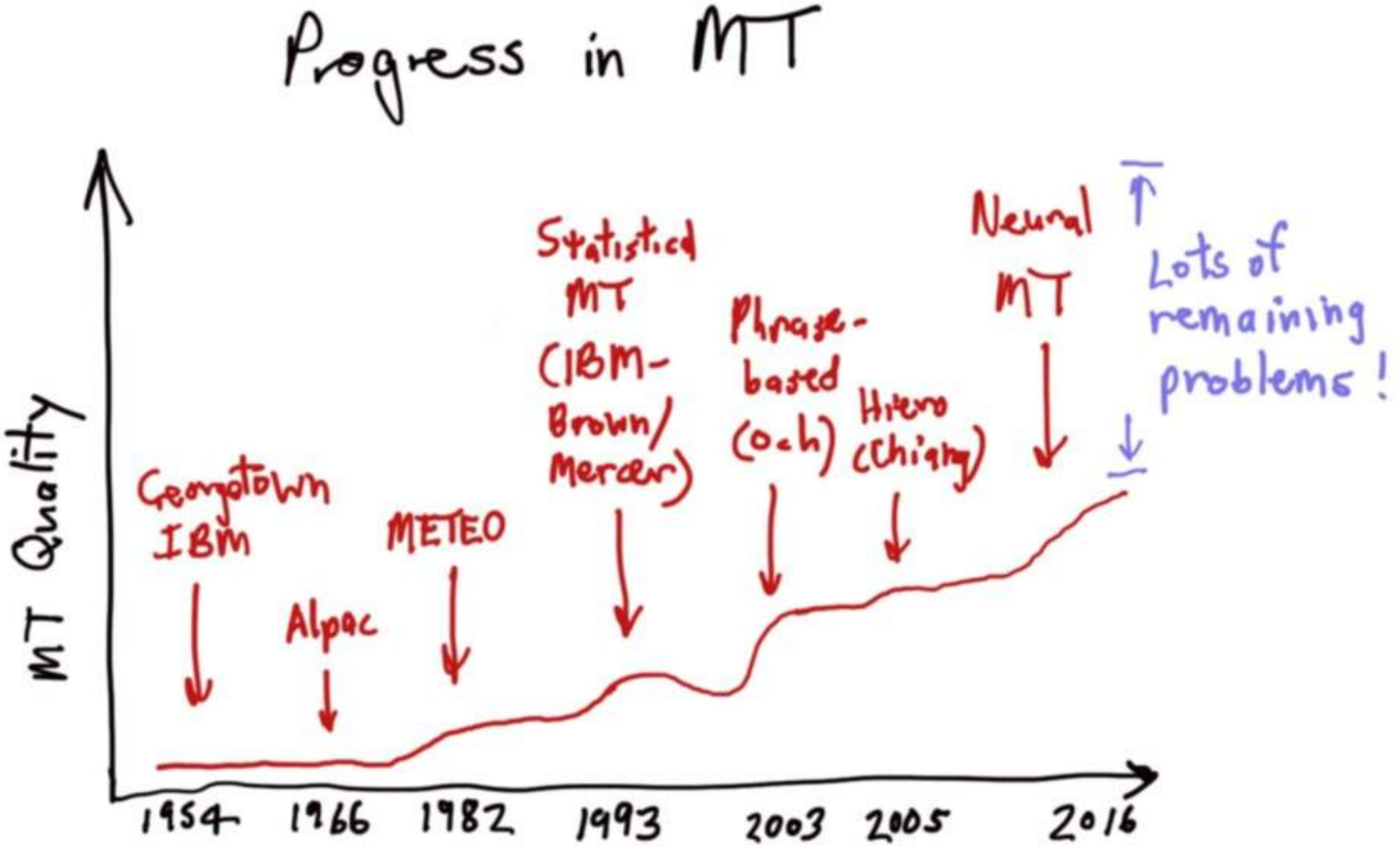
arxiv paper: https://arxiv.org/abs/1803.05567

Project Babel : a roadmap to Human Parity

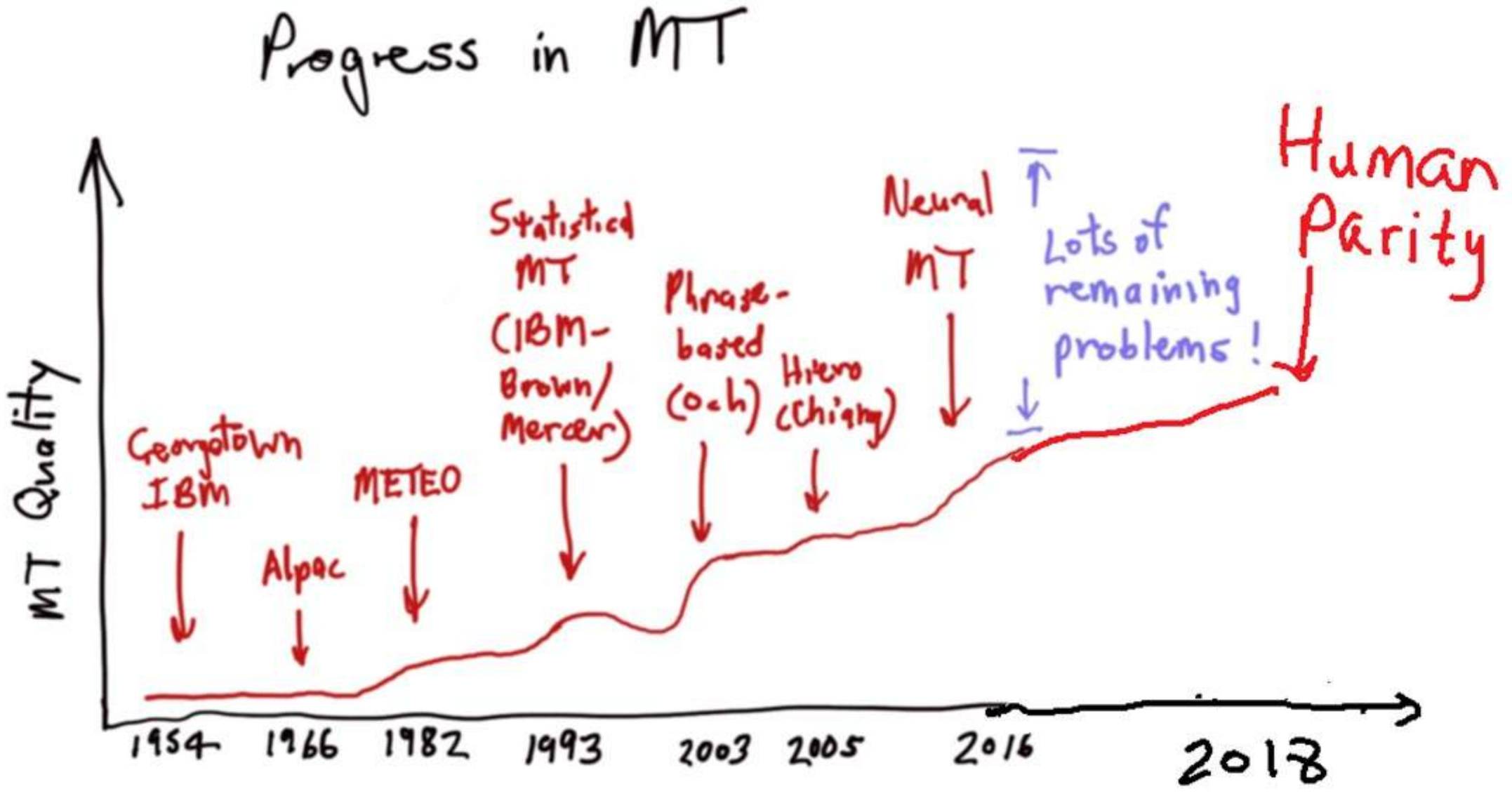
- Define new challenge for NMT research
- MT quality has improved a lot:
  - How far are we from human performance?
  - Fundamental question: How can we measure this?

- 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



Credit: Chris Manning 2016

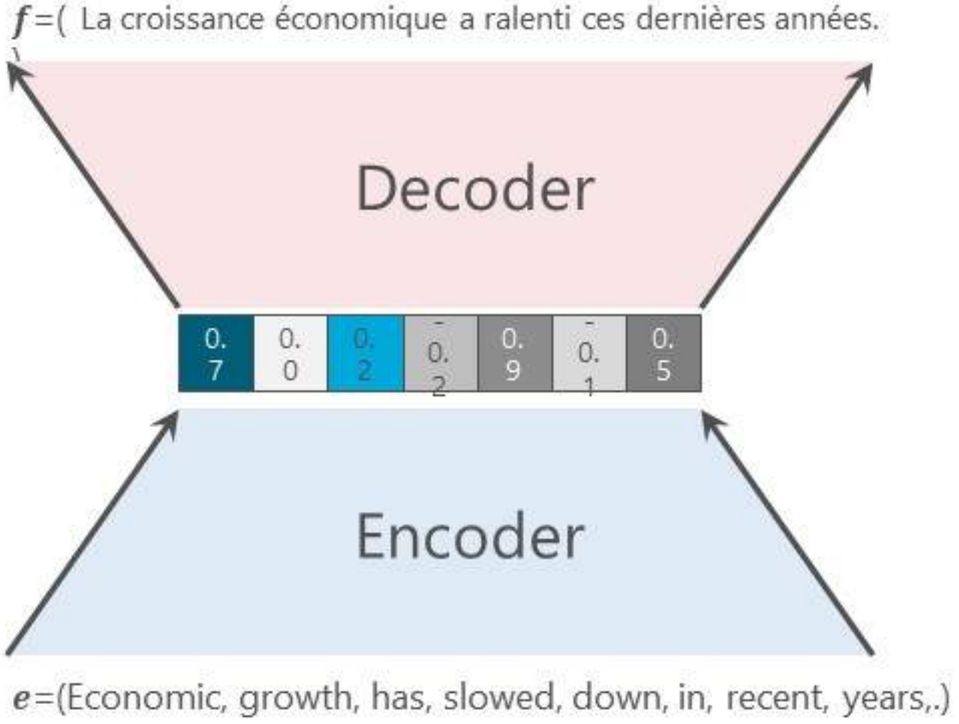


Edited from Chris Manning 2016

Outline

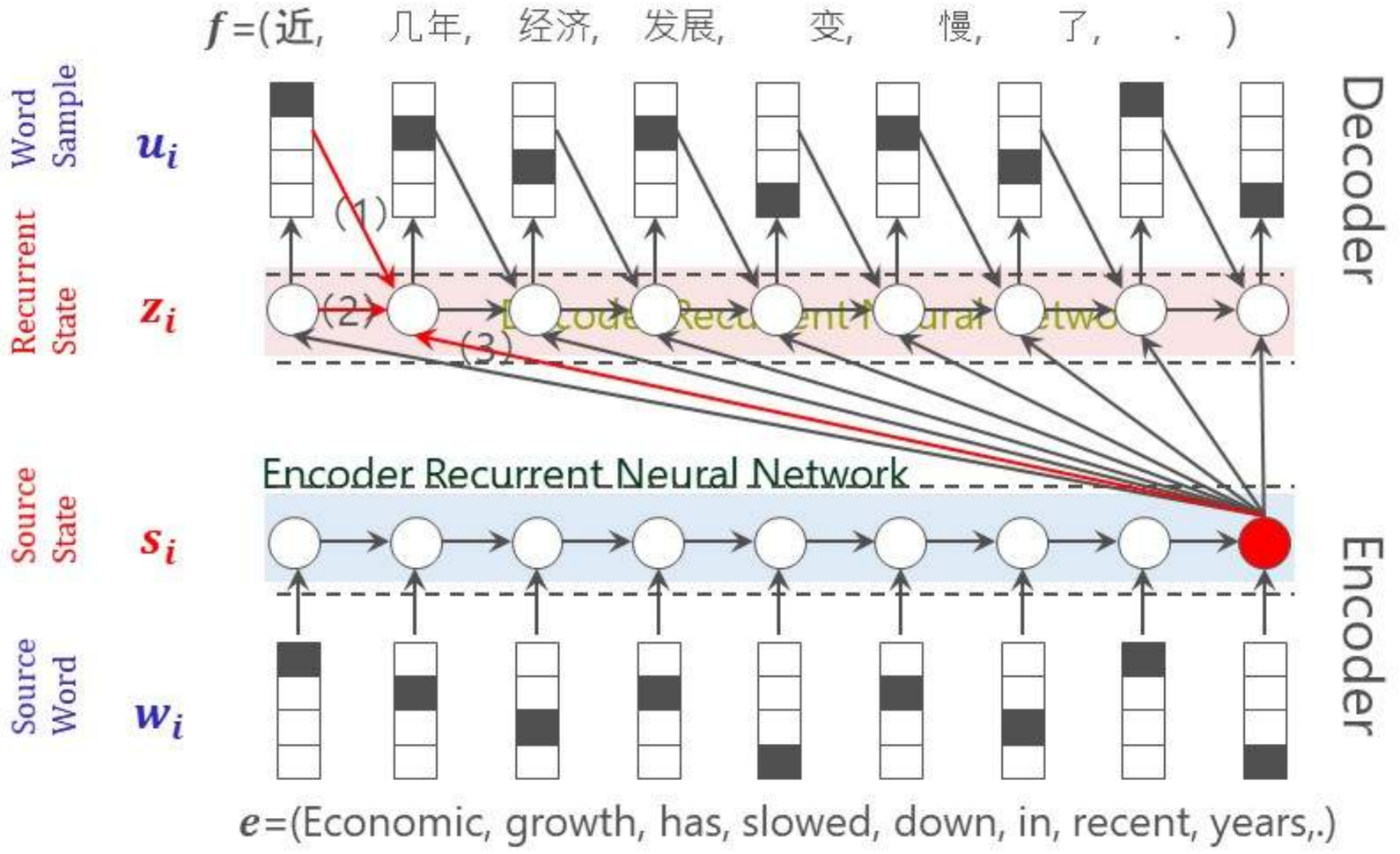
- NMT Overview
- Exploiting Dual Nature of Translation
  - Dual Learning
  - Joint Training
- Beyond left-to-right bias
  - Target Bidirectional Agreement
  - Deliberation Networks
- Noisy training data
- Systems Combination
- Experiments
- Human Evaluation

Encoder-Decoder for NMT



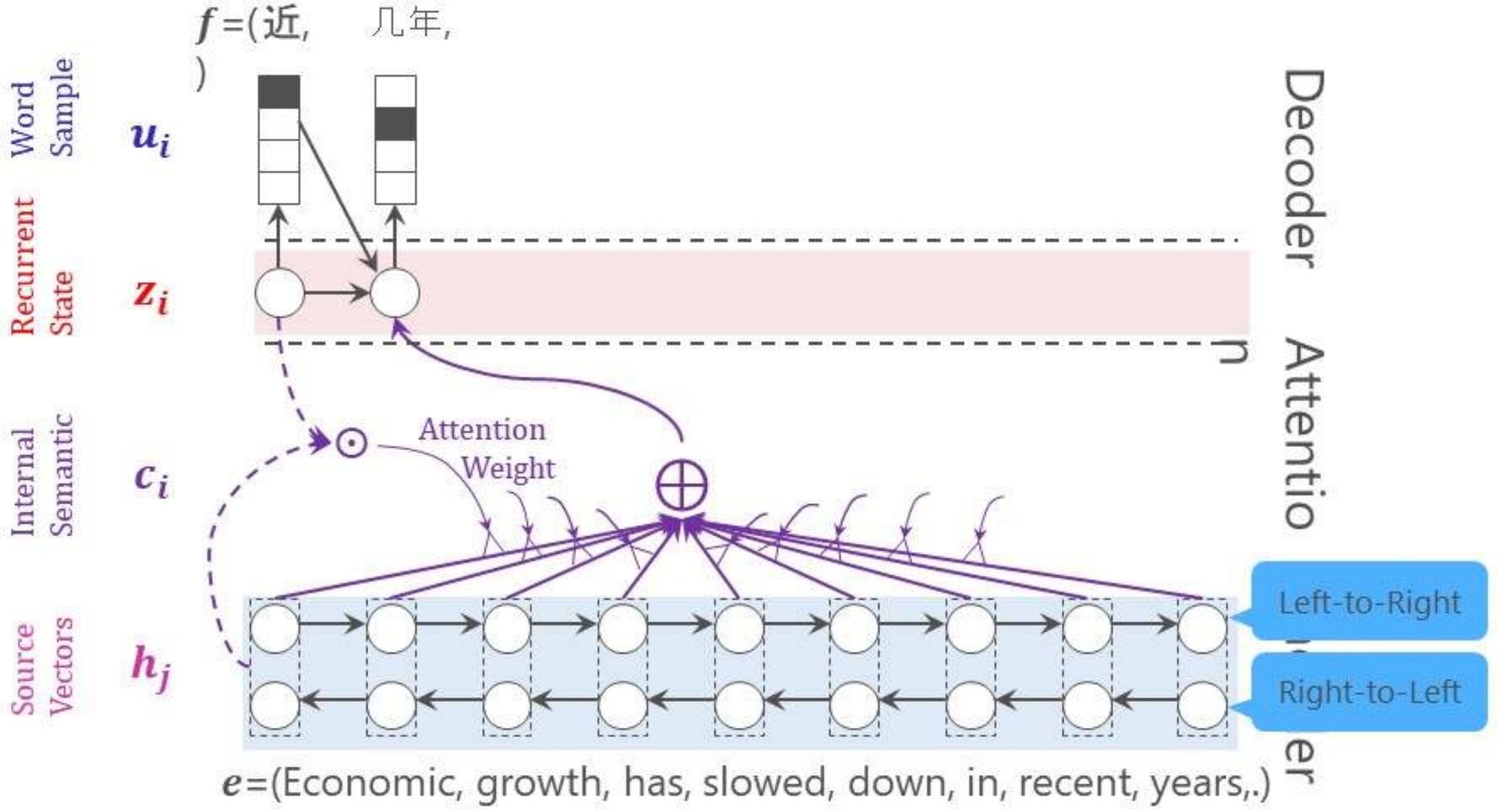
Sutskever et al., 2014, Cho et al., 2014

Encoder-Decoder for NMT



Sutskever et.al, NIPS 2014

Attention based Encoder-Decoder



Bahdanau et.al ICLR15

Seq2Seq with RNN models

- RNN
  - Very efficient representation for variable-length representations
  - LSTM/GRU gating helps in long-range error propagation
  - Attention helps in long-range dependencies
- Few limitations though:
  - Autoregressive models with sequential computation
  - Hard to parallelize
  - Hard to catch long-term dependencies in the sequential structure of the data
- CNN- based approaches like FairSeq(ConvS2S), ByteNet, WaveNet:
  - Easier to parallelize
  - Needs more layers to capture long-term dependencies

Attention is all you need: Transformer models

- Avoids RNN recurrence bottleneck
- Encoder:
  - Multiple layers of self-attention and Feed Forward Networks
- Decoder:
  - Self-Attention on decoder and attention on encoder outputs with Feed Forward Networks
- Multi-head Attention:
  - Model different information at different positions

Vaswani et.al NIPS 2017

Seq2Seq with RNN models

- RNN
  - Very efficient representation for variable-length representations
  - LSTM/GRU gating helps in long-range error propagation
  - Attention helps in long-range dependencies
- Few limitations though:
  - Autoregressive models with sequential computation
  - Hard to parallelize
  - Hard to catch long-term dependencies in the sequential structure of the data
- CNN- based approaches like FairSeq(ConvS2S), ByteNet, WaveNet:
  - Easier to parallelize
  - Needs more layers to capture long-term dependencies

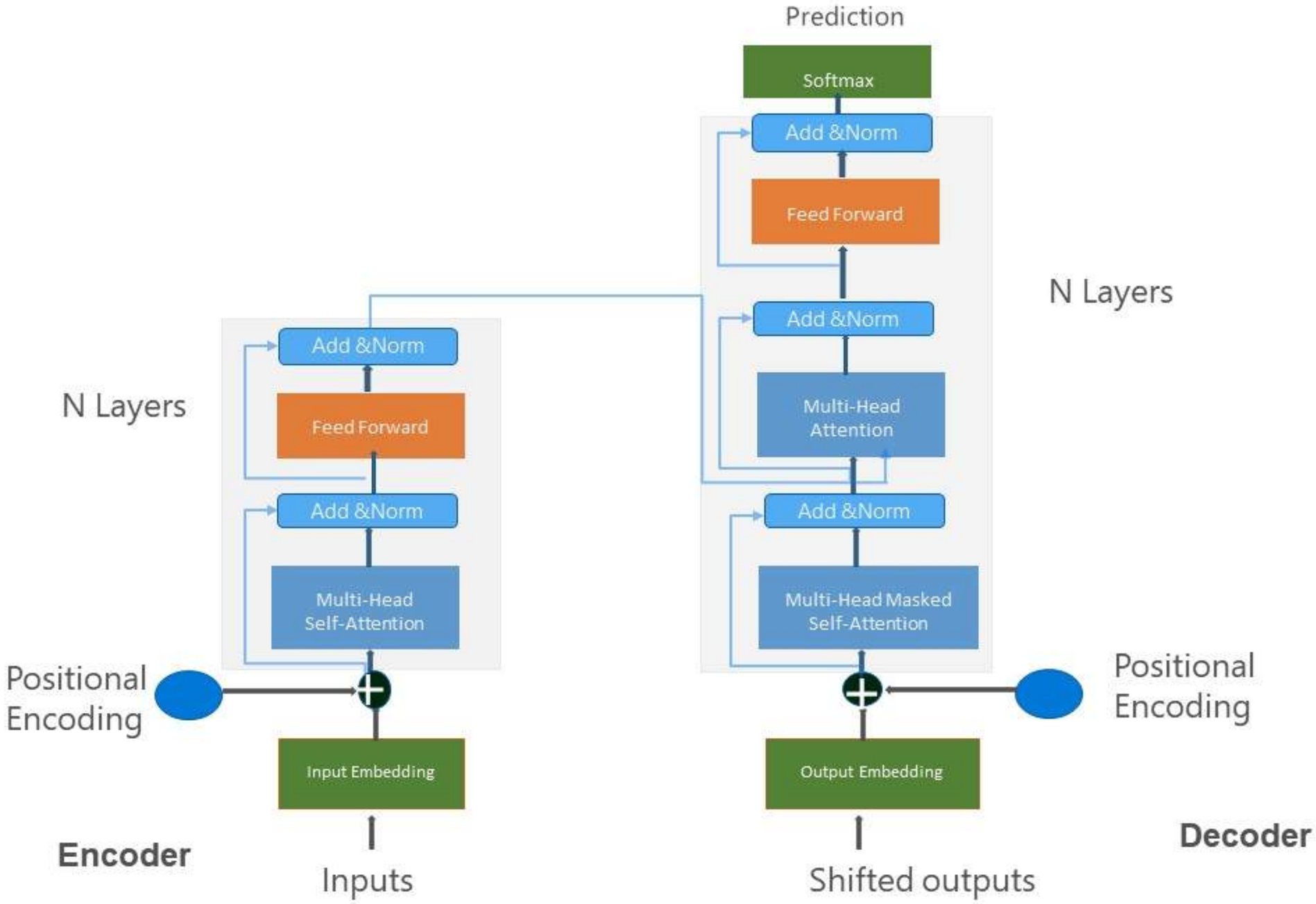
Attention is all you need: Transformer models

- Avoids RNN recurrence bottleneck
- Encoder:
  - Multiple layers of self-attention and Feed Forward Networks
- Decoder:
  - Self-Attention on decoder and attention on encoder outputs with Feed Forward Networks
- Multi-head Attention:
  - Model different information at different positions

Vaswani et.al NIPS 2017

Attention is all you need: Transformer models

Credit: Google Research blog

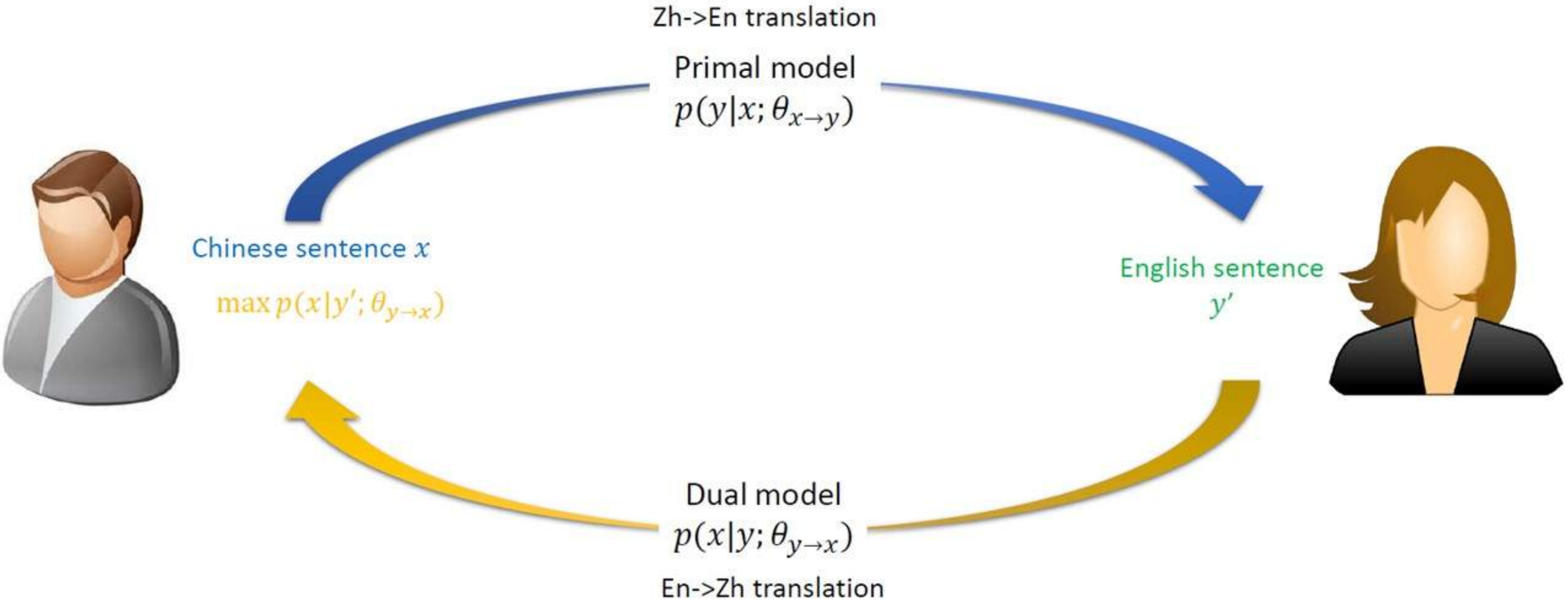

Prediction
Softmax
Add &Norm
Feed Forward
Add &Norm
Multi-Head Attention
Add &Norm
Multi-Head Masked Self-Attention
N Layers
Add &Norm
Feed Forward
Add &Norm
Multi-Head Self-Attention
N Layers
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Encoder
Decoder
Inputs
Shifted outputs

Outline

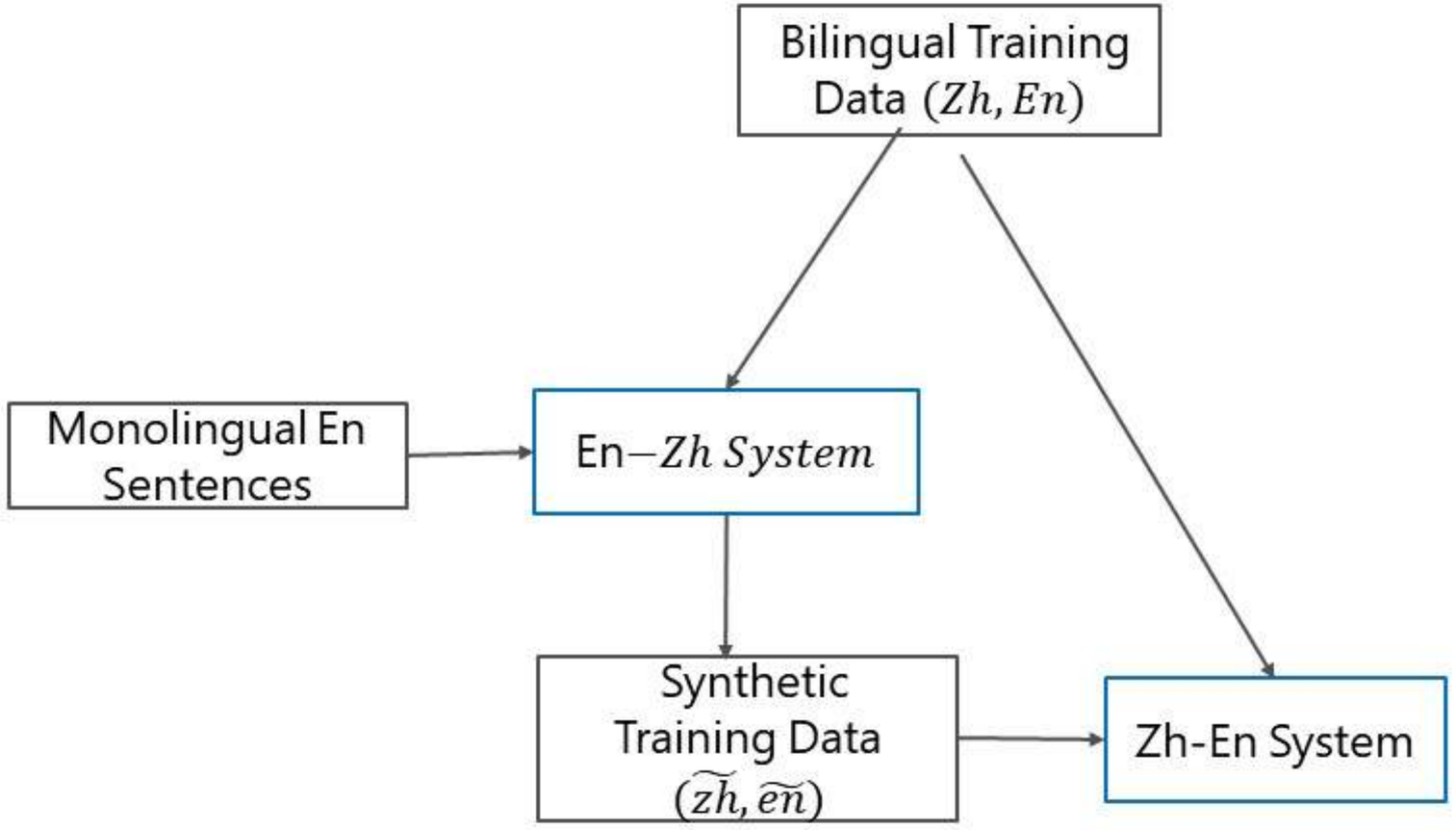
- NMT Overview
- Exploiting Dual Nature of Translation
  - Dual Learning
  - Joint Training
- Beyond left-to-right bias
  - Deliberation Networks
  - Target Bidirectional Agreement
- Noisy training data
- Systems Combination
- Experiments
- Human Evaluation

Exploiting Dual Nature of Translation

Use round-trip translation (Chinese → English → Chinese) to improve both systems simultaneously



Back Translation



MT Duality advantages over Back Translation

Back Translation

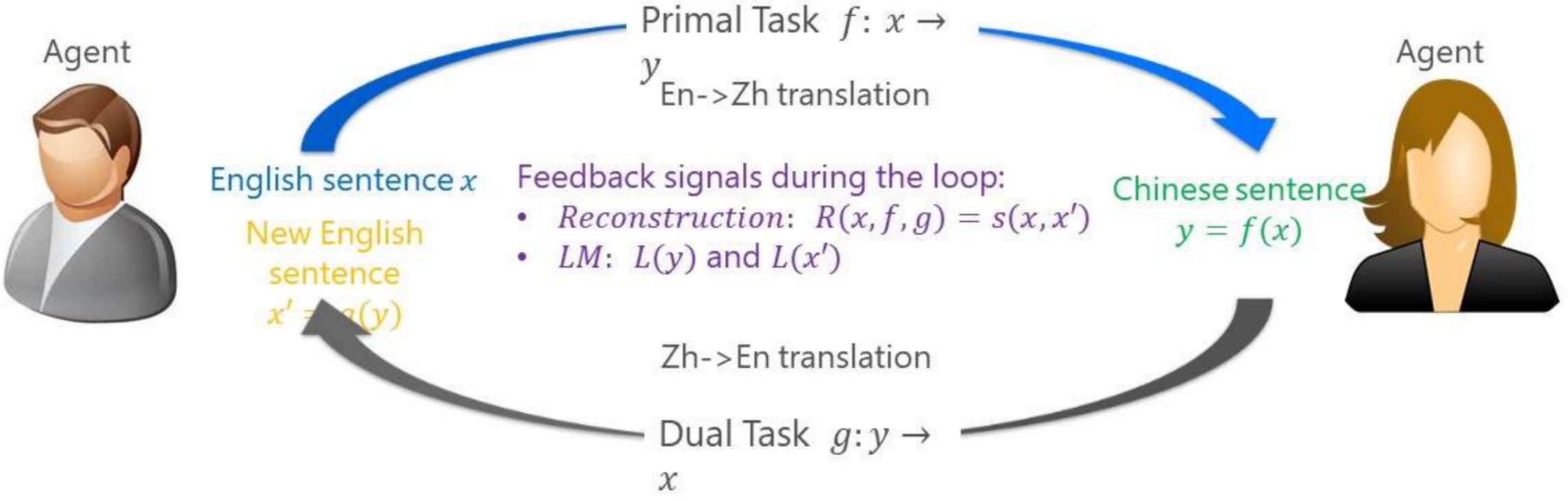
- Bad pseudo sentence pairs hurt the performance.
- Only the target monolingual data can be leveraged.
- Target to source translation model is not optimized.

MT Duality

- Bad pseudo sentence pairs cannot hurt the performance since it get partial credit
- Both Source and target monolingual data can be leveraged.
- Models in both directions can be optimized.
- Can be used in both semi-supervised and unsupervised setup

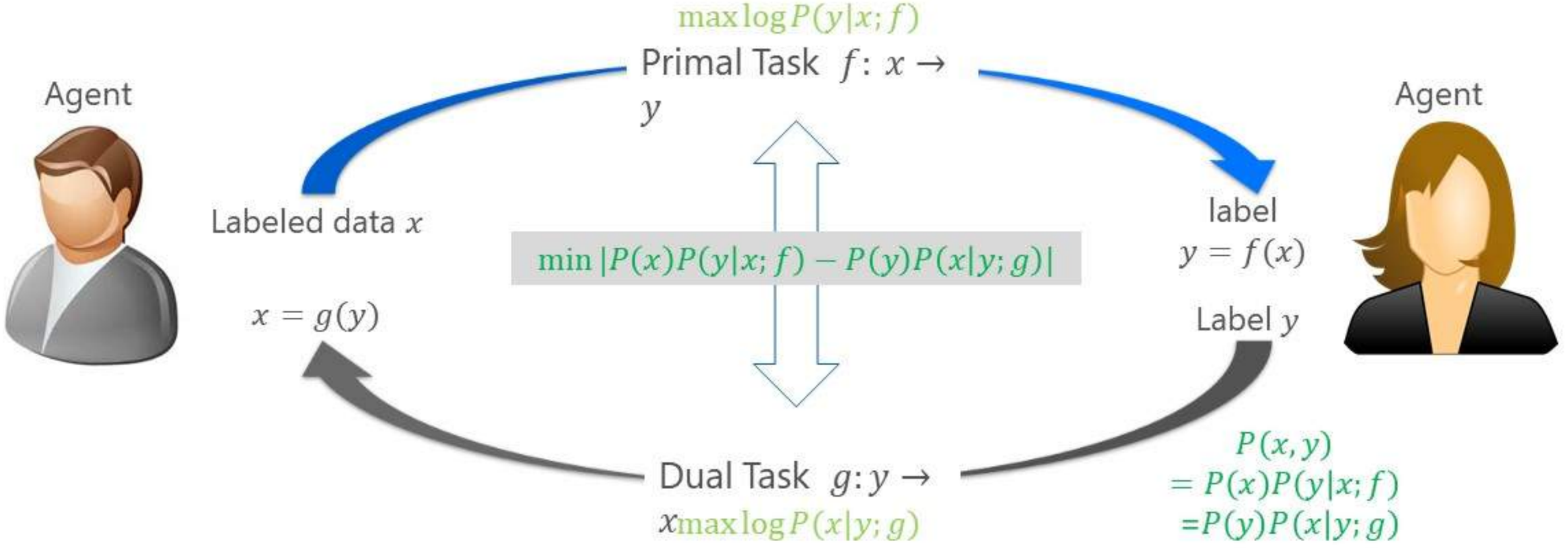
Dual Unsupervised Learning

(NIPS2016, ICML2017)



Dual Supervised Learning

(NIPS2016, ICML2017)

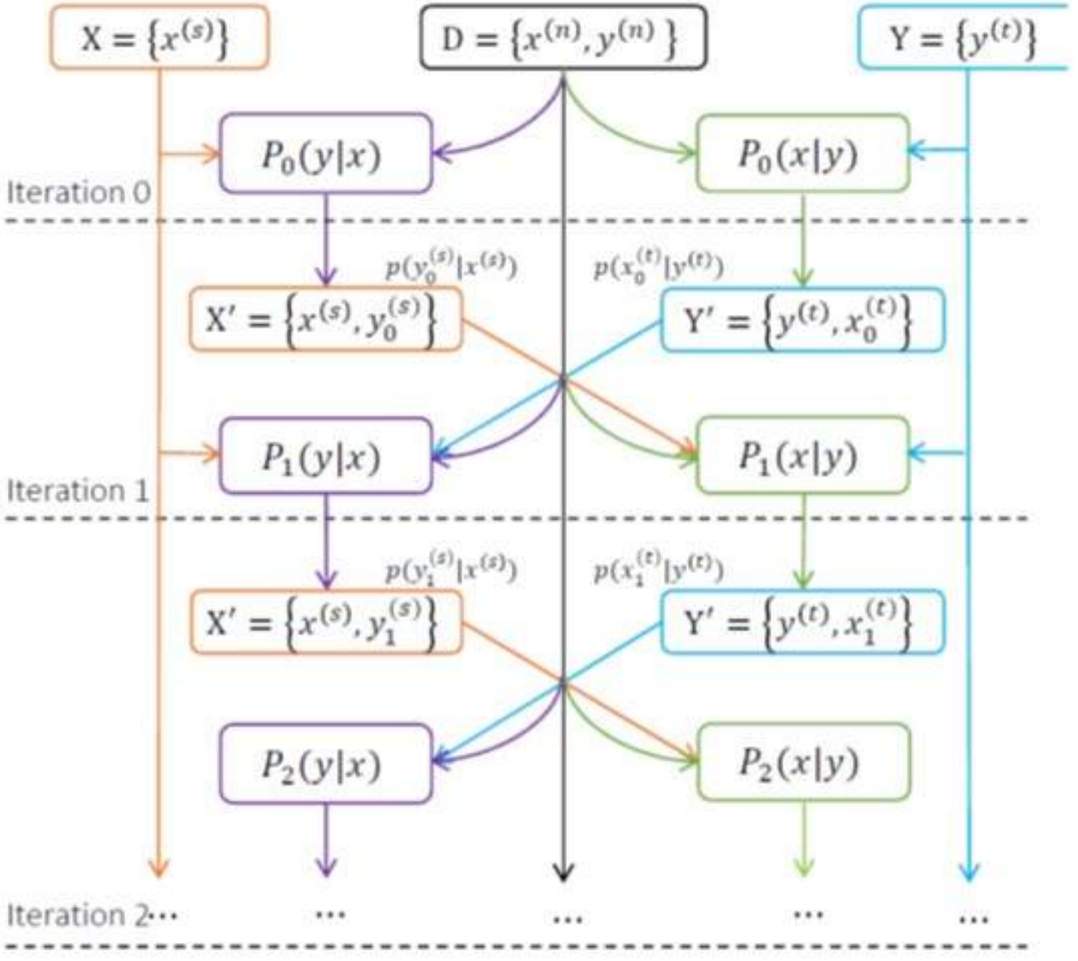


Feedback signals during the loop:

- $R(x, f, g) = |P(x)P(y|x; f) - P(y)P(x|y; g)|$: the gap between the joint probability $P(x, y)$ obtained in two directions

Joint Training (AAAI2018)

- Iteration 0: Pre-train two direction models $\mathbf{M}^{0}_{\mathbf{x}\to\mathbf{y}}$ and $\mathbf{M}^{0}_{\mathbf{y}\to\mathbf{x}}$ with bilingual data $\mathbf{D} = \{\mathbf{x}^{(\mathbf{n})}, \mathbf{y}^{(\mathbf{n})}\}$
- Iteration 1: Two NMT systems based on $\mathbf{P}^{0}_{\mathbf{x}\to\mathbf{y}}$ and $\mathbf{P}^{0}_{\mathbf{y}\to\mathbf{x}}$ are used to translate monolingual data $\mathbf{X} = \{\mathbf{x}^{(\mathbf{s})}\}$ and $\mathbf{Y} = \{\mathbf{y}^{(\mathbf{t})}\}$; two synthetic training data sets $\mathbf{X}'$ and $\mathbf{Y}'$ combined with bilingual data $\mathbf{D}$ are used to train $\mathbf{P}^{1}_{\mathbf{x}\to\mathbf{y}}$ and $\mathbf{P}^{1}_{\mathbf{y}\to\mathbf{x}}$ respectively.
- Iteration 2: Repeat the above process
- ....

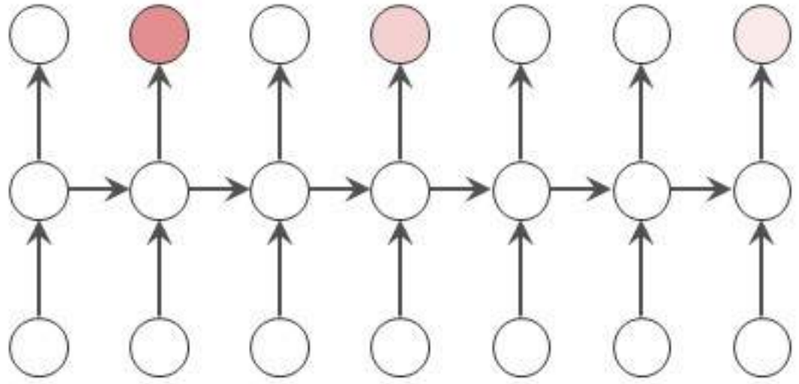


Outline

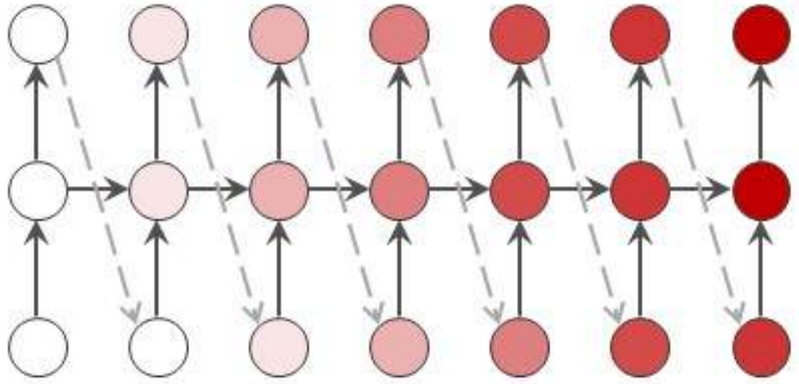
- NMT Overview
- Exploiting Dual Nature of Translation
  - Dual Learning
  - Joint Training
- Beyond left-to-right bias
  - Deliberation Networks
  - Target Bidirectional Agreement
- Noisy training data
- Systems Combination
- Experiments
- Human Evaluation

Motivation: Exposure Bias Problem

- NMT model is only trained with golden bilingual corpus
- Translation sentence is auto-regressively generated word by word
- Previous errors will mislead the generation of the subsequences
- Errors will be quickly amplified



Training



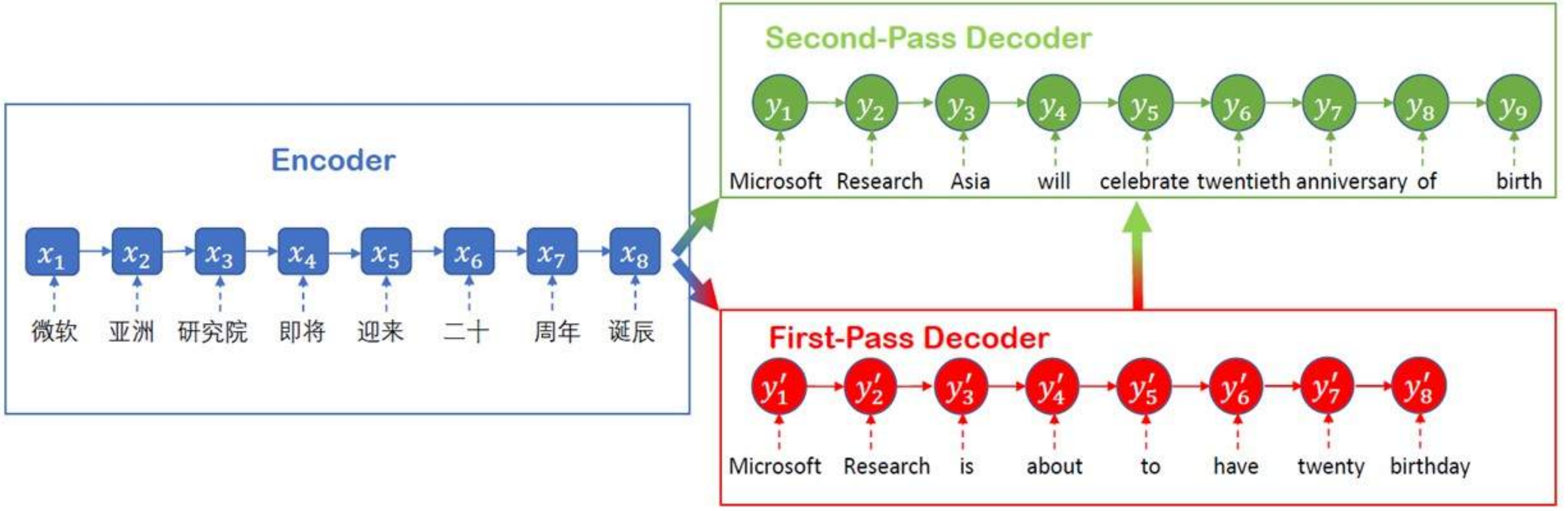
Decoding

Multi-pass decoding

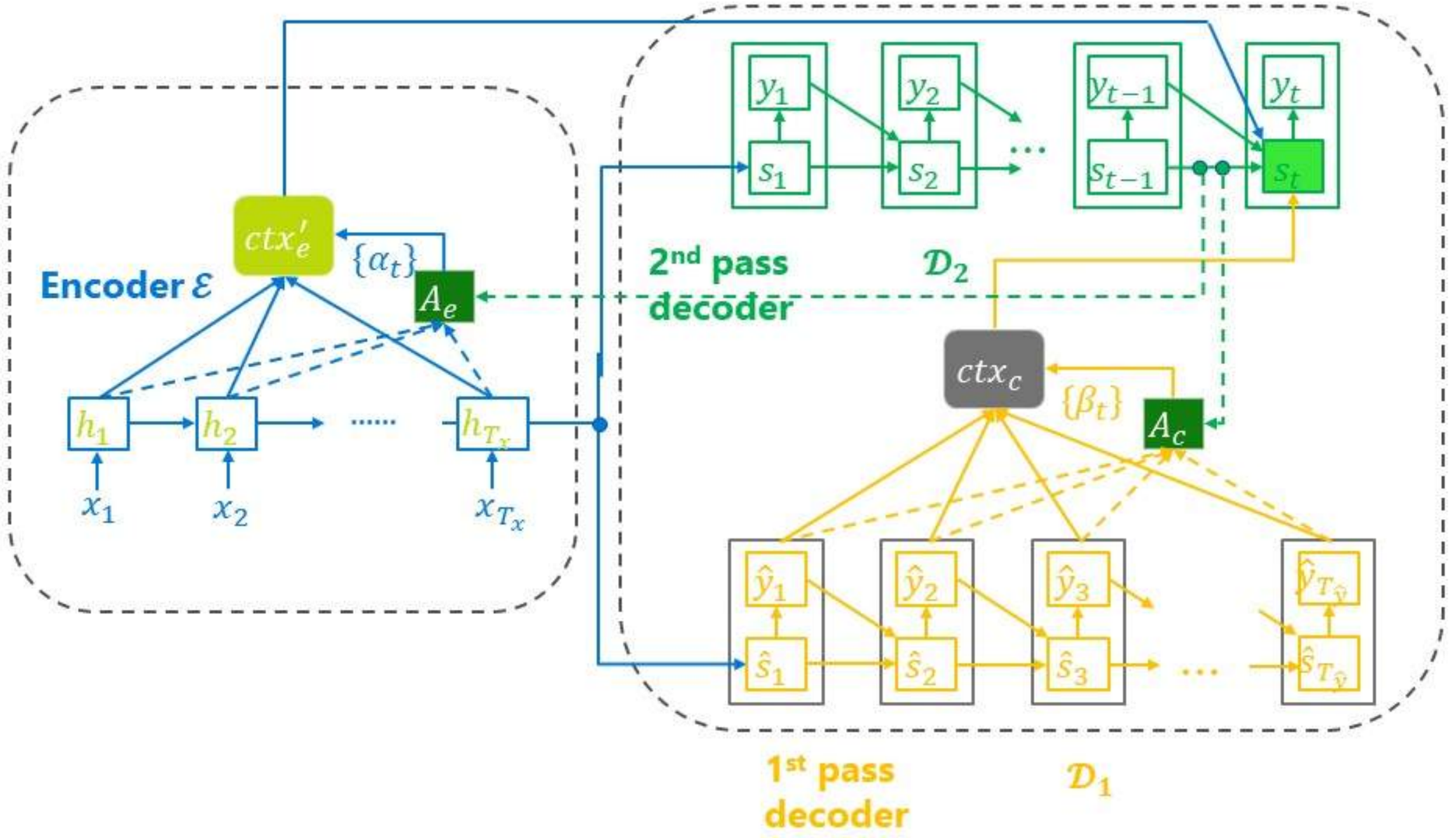
- Current NMT models decode from left to right in one pass
- To generate the word y_t, only the previous (already generated) words $y_{<t}$ are used
- In human cognitive processes, global information, including both historical words $y_{<t}$ and possible future words $y_{>t}$ are used
- While reading a document/paragraph/sentence, our understanding of one word depends on its context, including both the words preceding and after it
- When we write a document/article, we first create a draft and then polish it based on global understanding of the whole draft

Deliberation networks (NIPS2017)

Multi-pass translation: Translate → Refine → Translate again



Deliberation Nets



Beyond the Left-to-Right Bias

During sequential left-to-right generation of translation earlier errors are amplified.

Seeking agreement between left-to-right and right-to-left translation at training time

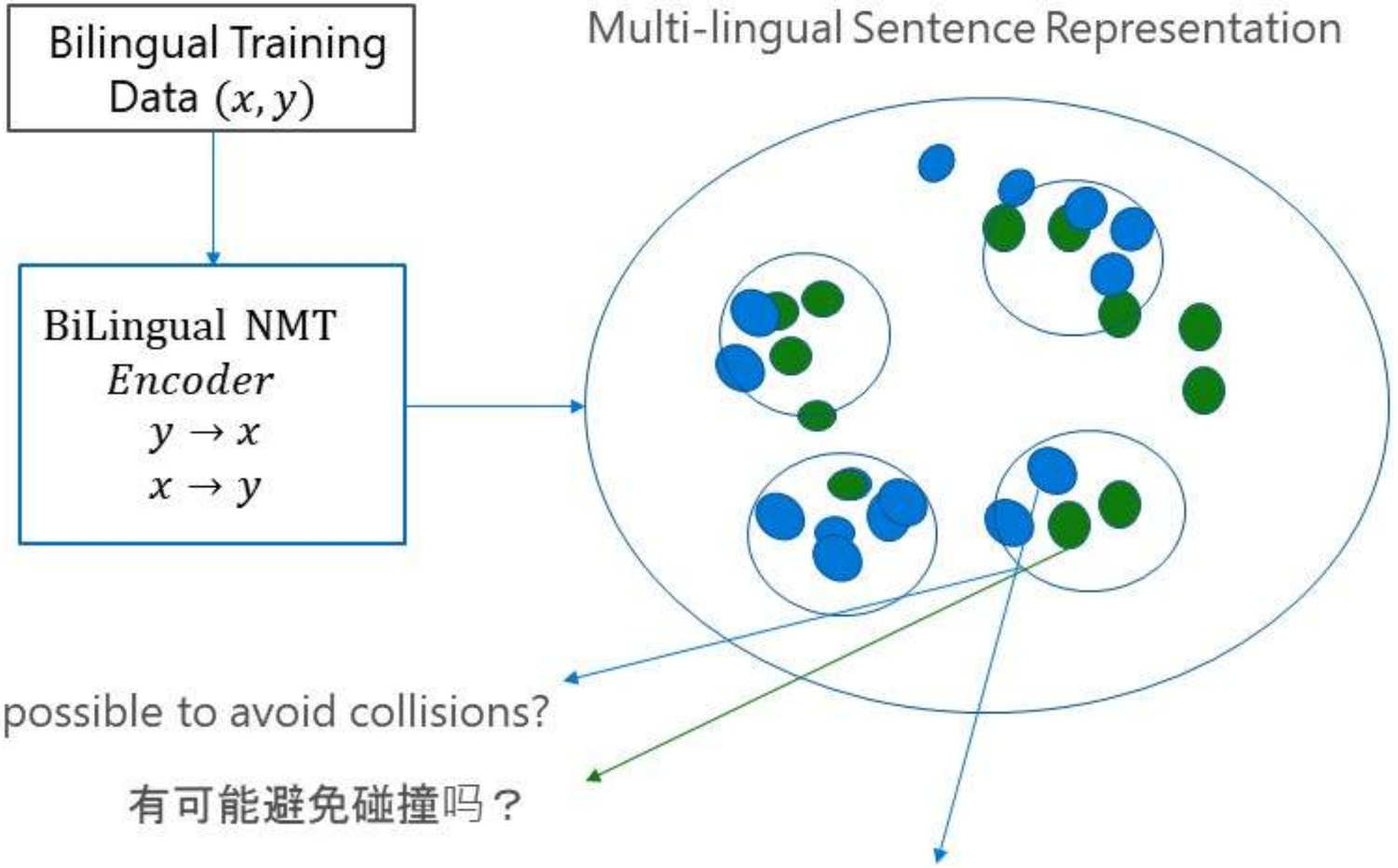
Iterative training of left-to-right and right-to-left models boost each other

Outline

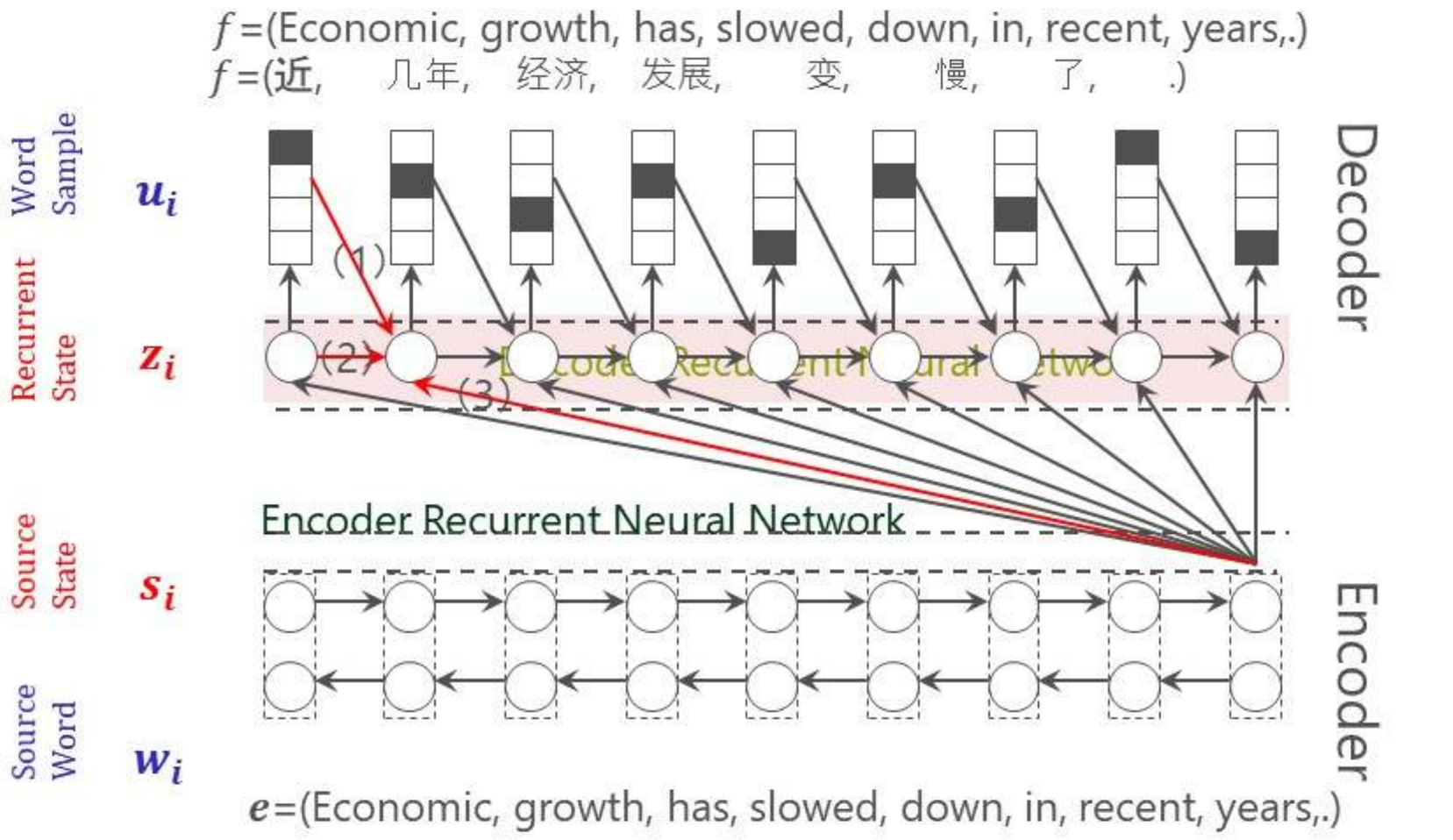
- NMT Overview
- Exploiting Dual Nature of Translation
  - Dual Learning
  - Joint Training
- Beyond left-to-right bias
  - Target Bidirectional Agreement
  - Deliberation Networks
- Noisy training data
- Systems Combination
- Experiments
- Human Evaluation

Data Selection and Filtering

- Train a multi-lingual sentence representation (SentVec)
- Semantically similar sentences in either language are close by in the space
- Use distance in the space to select sentence pairs where English & Chinese are close, indicating the data is of good quality
- Also used Cross-Entropy Difference (Moore-Lewis 2010) to prefilter data for backtranslation and dual learning

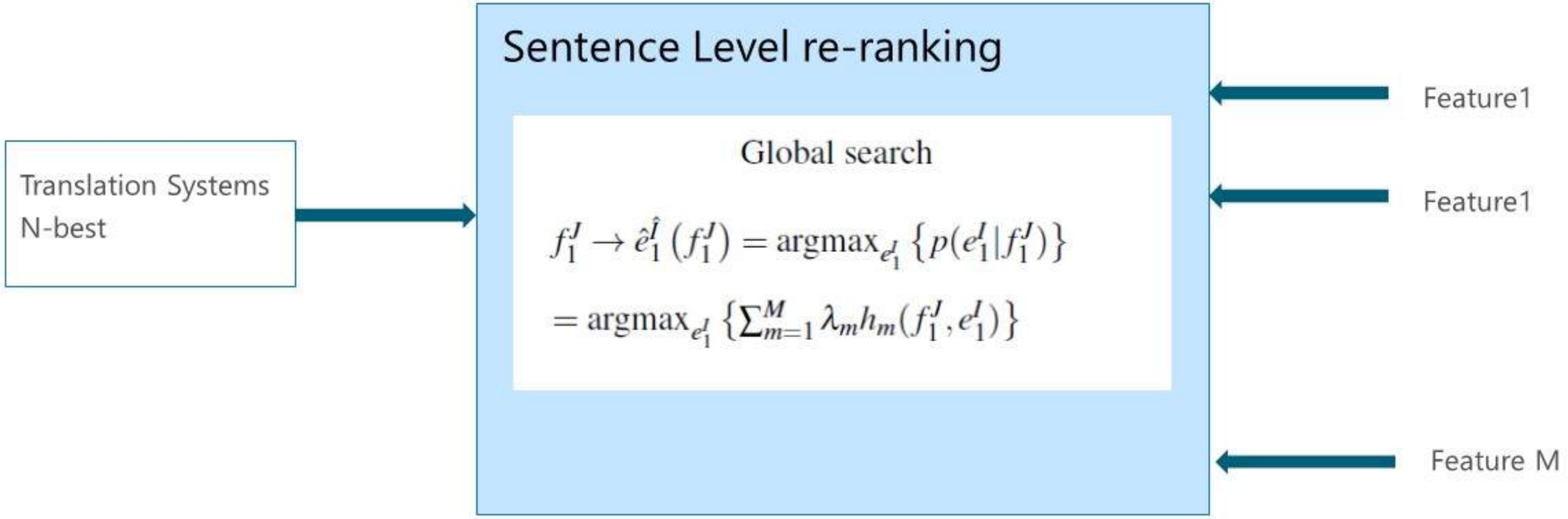


Bi-Lingual Encoder-Decoder System

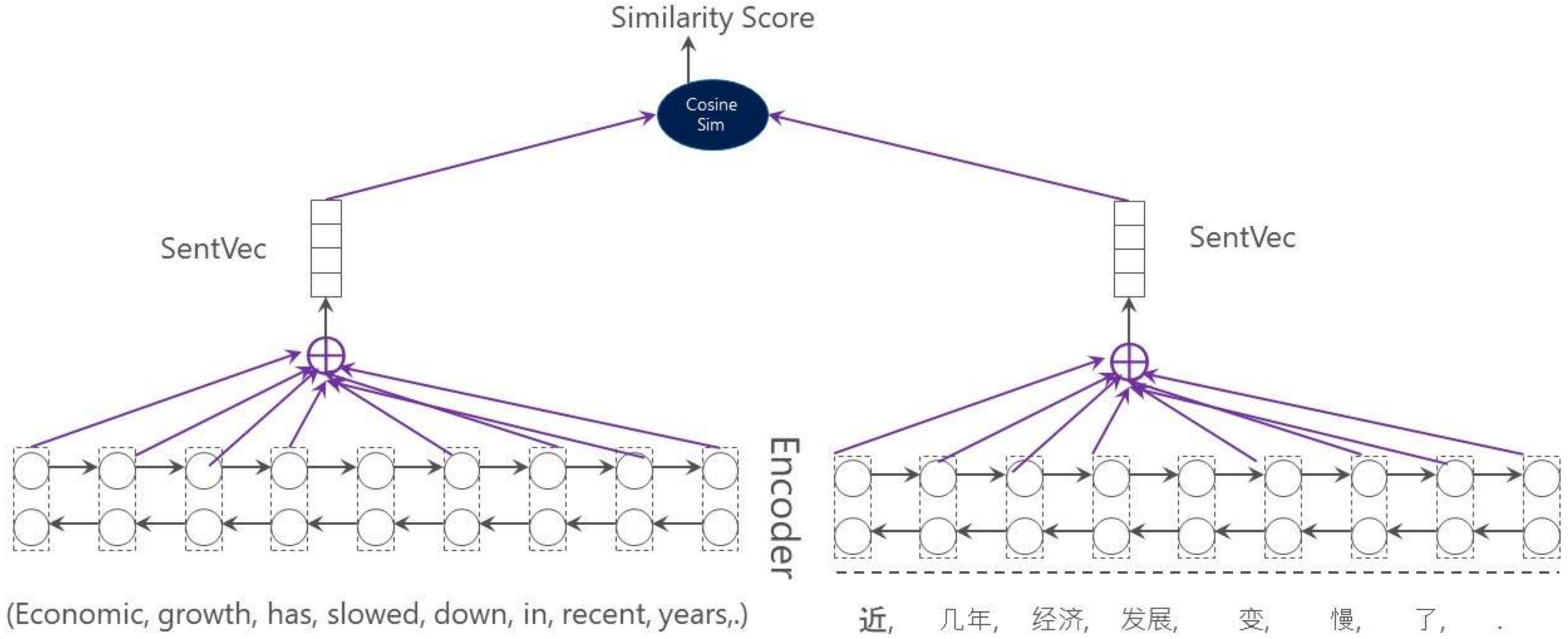


System combination & reranking

- Combine multiple systems: different types of expertise, cancel out each other's mistakes
- Re-rank n-best list: Use whole-sentence features

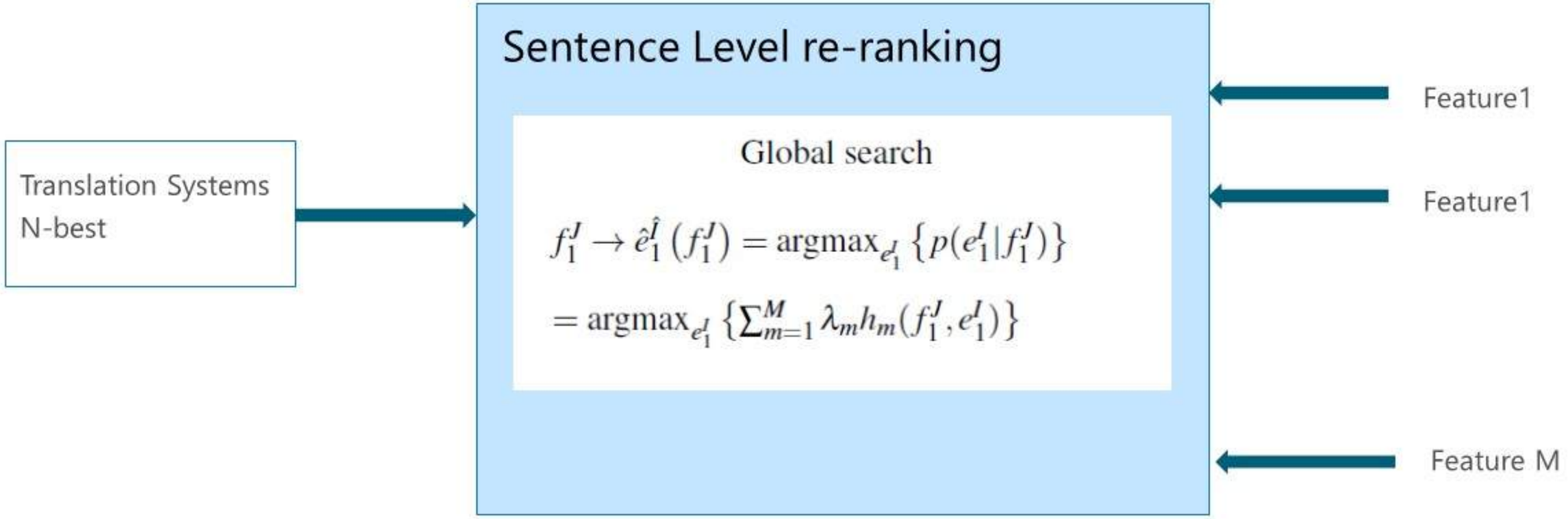


SentVec Similarity Model



System combination & reranking

- Combine multiple systems: different types of expertise, cancel out each other's mistakes
- Re-rank n-best list: Use whole-sentence features



Systems Combination Features

- SYS_{Score}: Original System Score and identifier.
- LM_{Score}: 5-gram language model trained on English news crawled data of 2015 and 2016.
- $R2L_{score}$: R2L system re-scoring. A system trained on Chinese source and reversed English target; the system is used to score each hypothesis.
- $E2Z_{score}$: English-to-Chinese system re-scoring. A system trained on English to Chinese is used to score each hypothesis. .
- ST_{SV} : Cross-lingual sentence similarity between source and the hypothesis as described in Section 3.5.
- $R2L_{SV}$: R2L sentence vector similarity: the best hypothesis from the R2L system is compared to each n-best hypothesis and used to generate a sentence similarity score based on sentence vector as above.
- $E2Z_{SV}$: Back Composition sentence vector similarity. A round trip translation is done for each n-best hypothesis to translate it back to Chinese. Then we use sentence vector similarity to measure the similarity between the original source and the recomposed source.

Experimental Results

- Automatic (BLEU) evaluation results on the WMT 2017 Chinese-English test set

SystemID	Settings	BLEU
Sogou	WMT 2017 best result [42]	26.40
Base	Transformer Baseline	24.2
BT	+Back Translation	25.57
DL	BT + Dual Learning	26.51
DLDN	BT + Dual Learning + Deliberation Nets	27.40
DLDN2	DLDN without first decoder reranking	27.20
DLDN3	BT+ Dual Learning + R2L sampling	26.88
DLDN4	BT+ Dual Learning + Bi-NMT	27.16
AR	BT + Agreement Regularization	26.91
ARJT	BT + Agreement Regularization + Joint Training	27.38
ARJT2	ARJT + dropout=0.1	27.19
ARJT3	ARJT + dropout=0.05	27.07
ARJT4	ARJT + dropout=0.01	26.98

Data Selections Experiments

SystemID	Settings	BLEU
Base	Transformer Baseline	24.2
BT	+Back Translation	25.57
Base8K	BT + 8K d_{ff}	26.13
CED1	Base8K + 35M CED + dropout=0.1	26.68
CED2	Base8K + 50M CED + dropout=0.1	26.61
SV1	Base8K + 35M + dropout=0.1	27.60
SV2	Base8K + 50M + dropout=0.1	27.45
SV3	Base8K + 35M + dropout=0.2	27.67
SV4	Base8K + 50M + dropout=0.2	27.49

Table 2: Evaluation Data selection results on the WMT 2017 Chinese-English test set

Systems Combination

SystemID	Settings	BLEU
Combo-1	SV1, SV2, SV3	27.84
Combo-2	DLDN2, DLDN3, DLDN4	27.92
Combo-3	ARJT2, ARJT3, ARJT4 + 3 identical systems with different initialization	27.82
Combo-4	SV1, SV2, SV3, ARJT1, ARJT2, ARJT3, DLDN2, DLDN3, DLDN4	28.46
Combo-5	SV1, SV2, SV3, ARJT2, DLDN2, DLDN4	28.32
Combo-6	SV1, SV2, SV4, ARJT2, ARJT3, ARJT4, DLDN2, DLDN3, DLDN4	28.42

Table 3: System combination results on the WMT 2017 Chinese-English test set

NMT Research Challenges

- The recent progress in Sequence-to-Sequence modeling from the whole research community has paved the road for this achievement.
- We have introduced a number of approaches that helped in reaching human parity for Chinese to English news translation.
- Many research areas worth investigating:
  - OOV/rare word problem and Named Entities
  - Low resource languages
  - Domain/topic adaptation
  - Document-level context translation
  - Multi-modality: text, speech, video, emotions, gaze, gender, context
  - Additional neural infrastructures (i.e.: Fusion of Transformer, RNN and CNN.)
  - Modeling better reasoning and language understanding during translation.

Human Evaluation

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

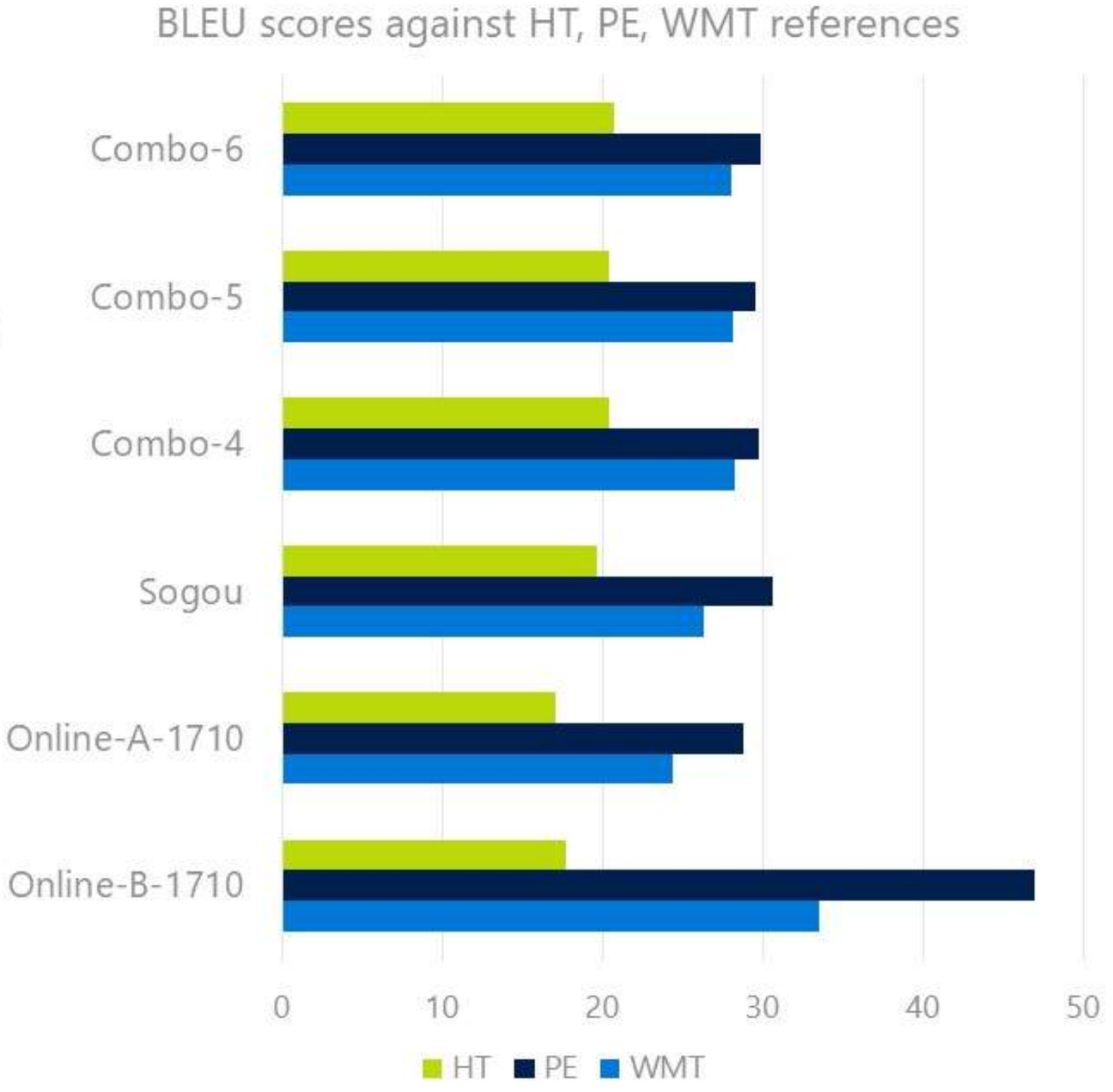
- Created two new references:
  - PE = post-edited / crowd-sourced
  - HT = human translation from scratch

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

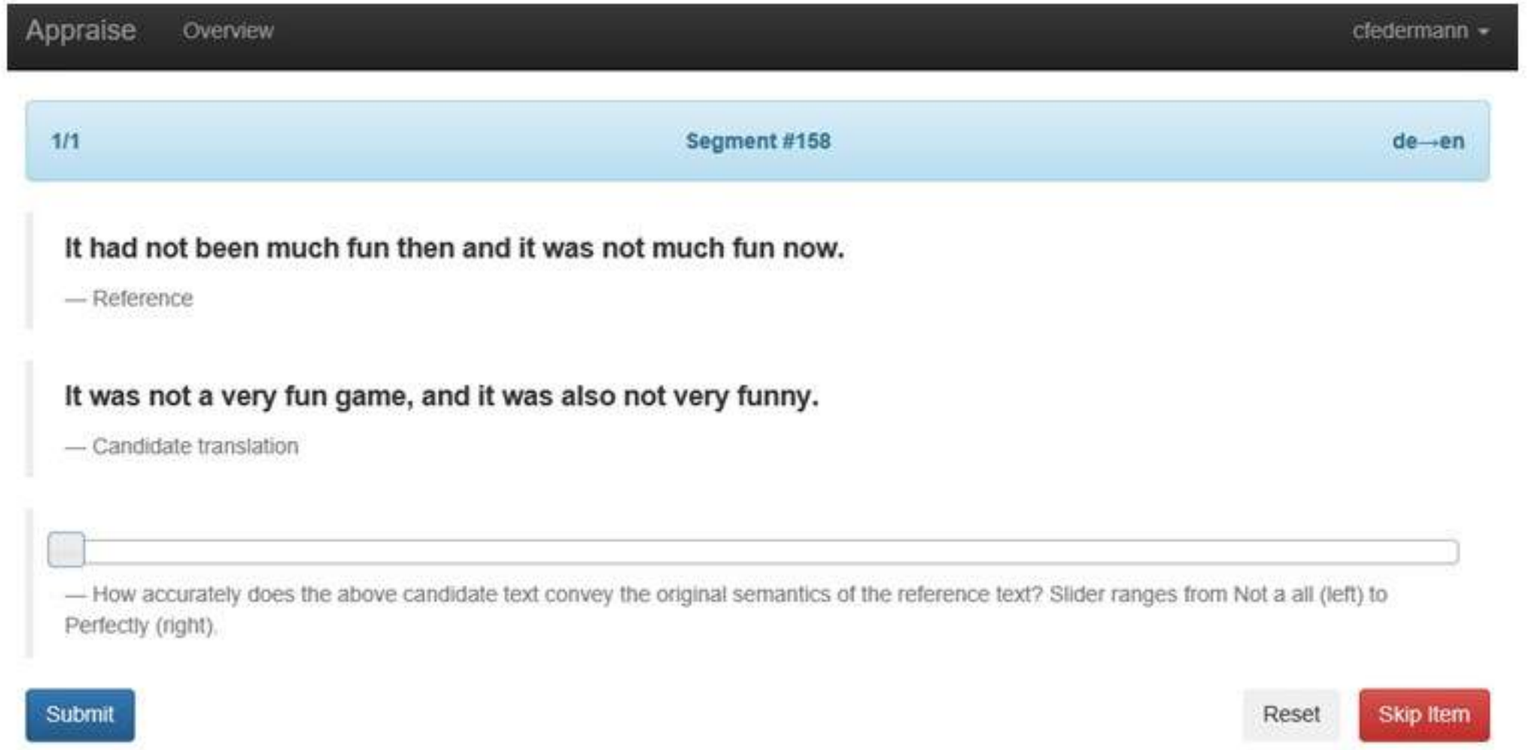
Direct assessment

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

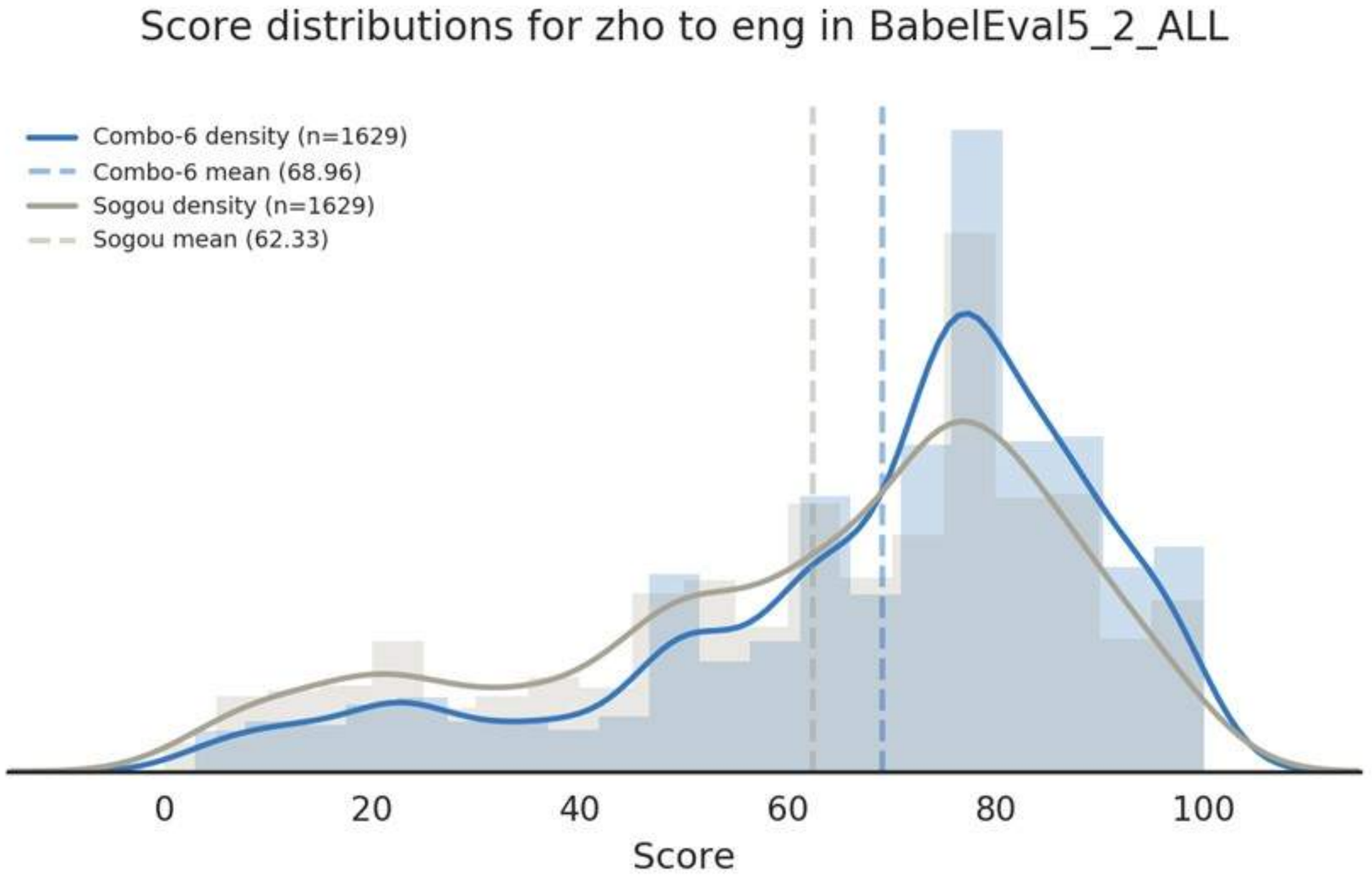
From

Result clusters for all systems

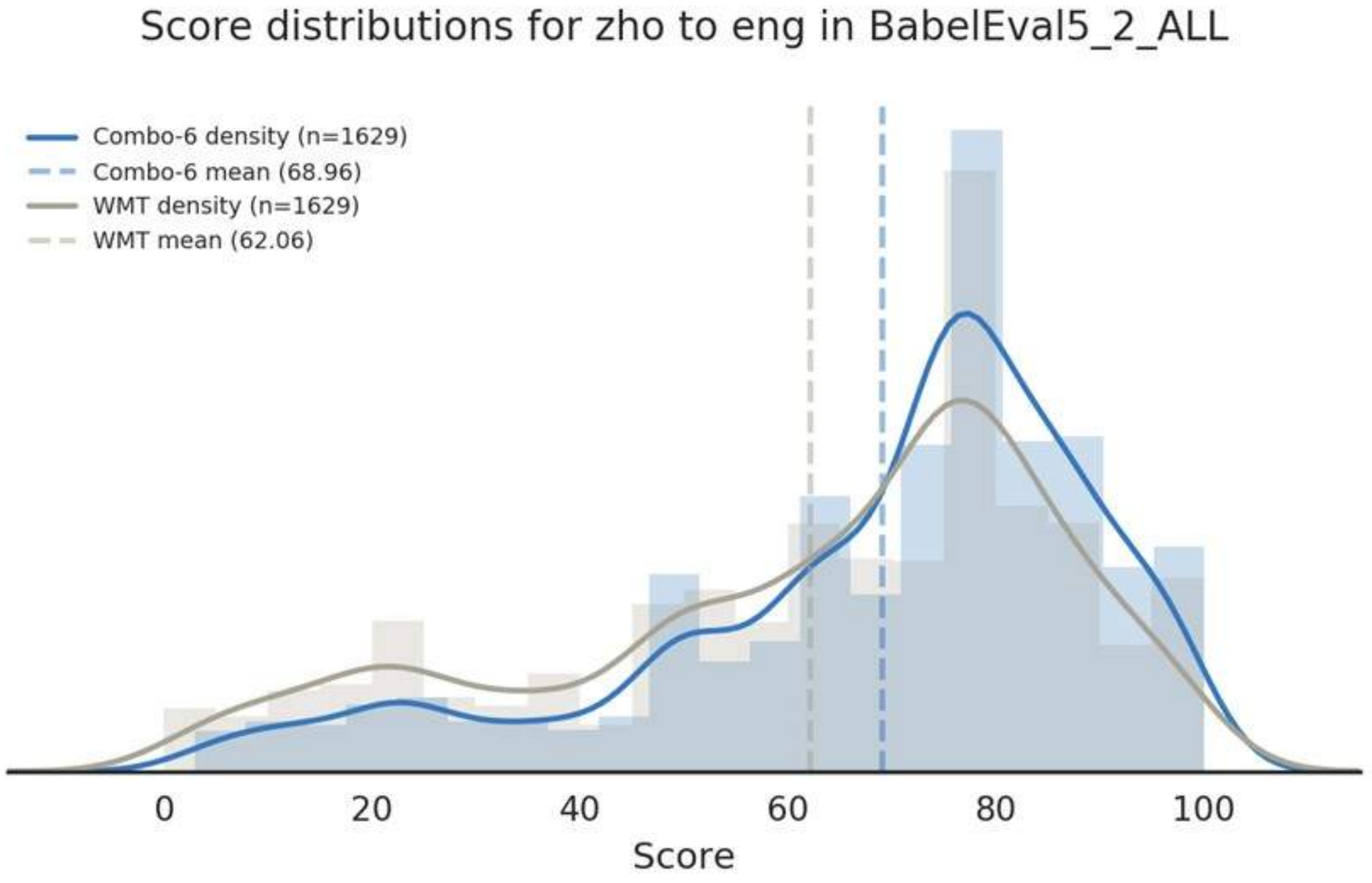
To

Pairwise density representation

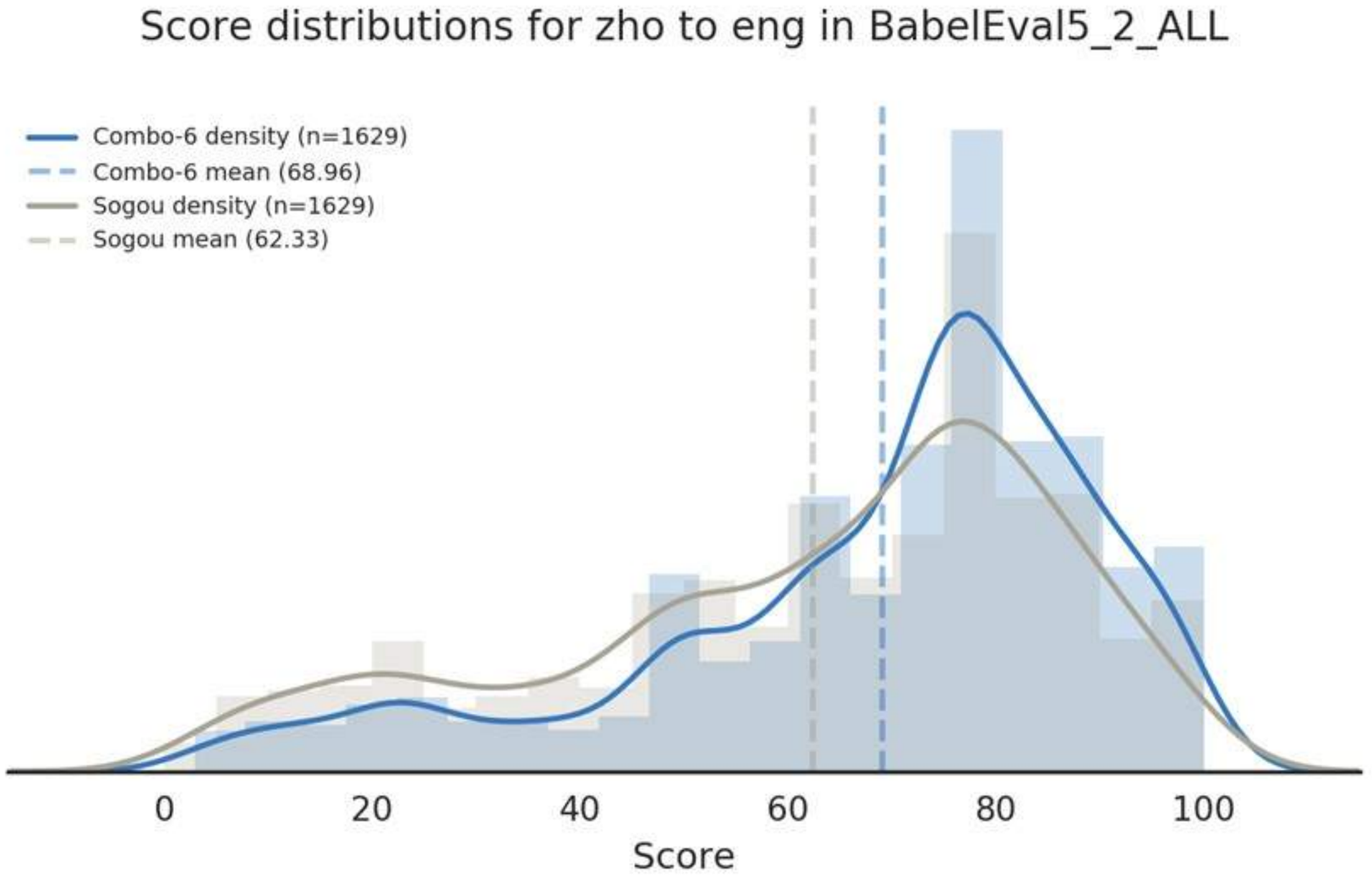
Combo-6 vs Sogou



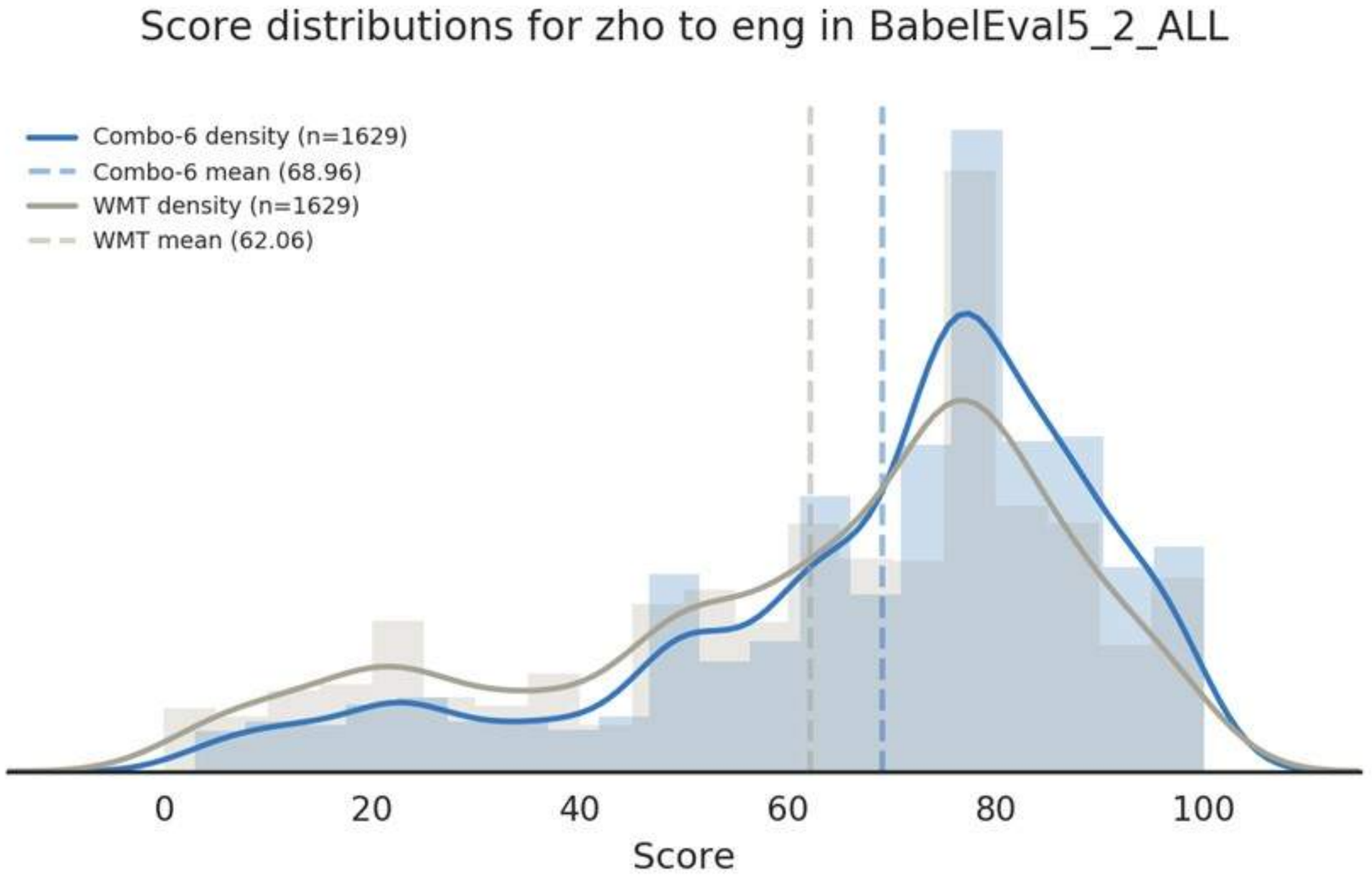
Combo-6 vs WMT



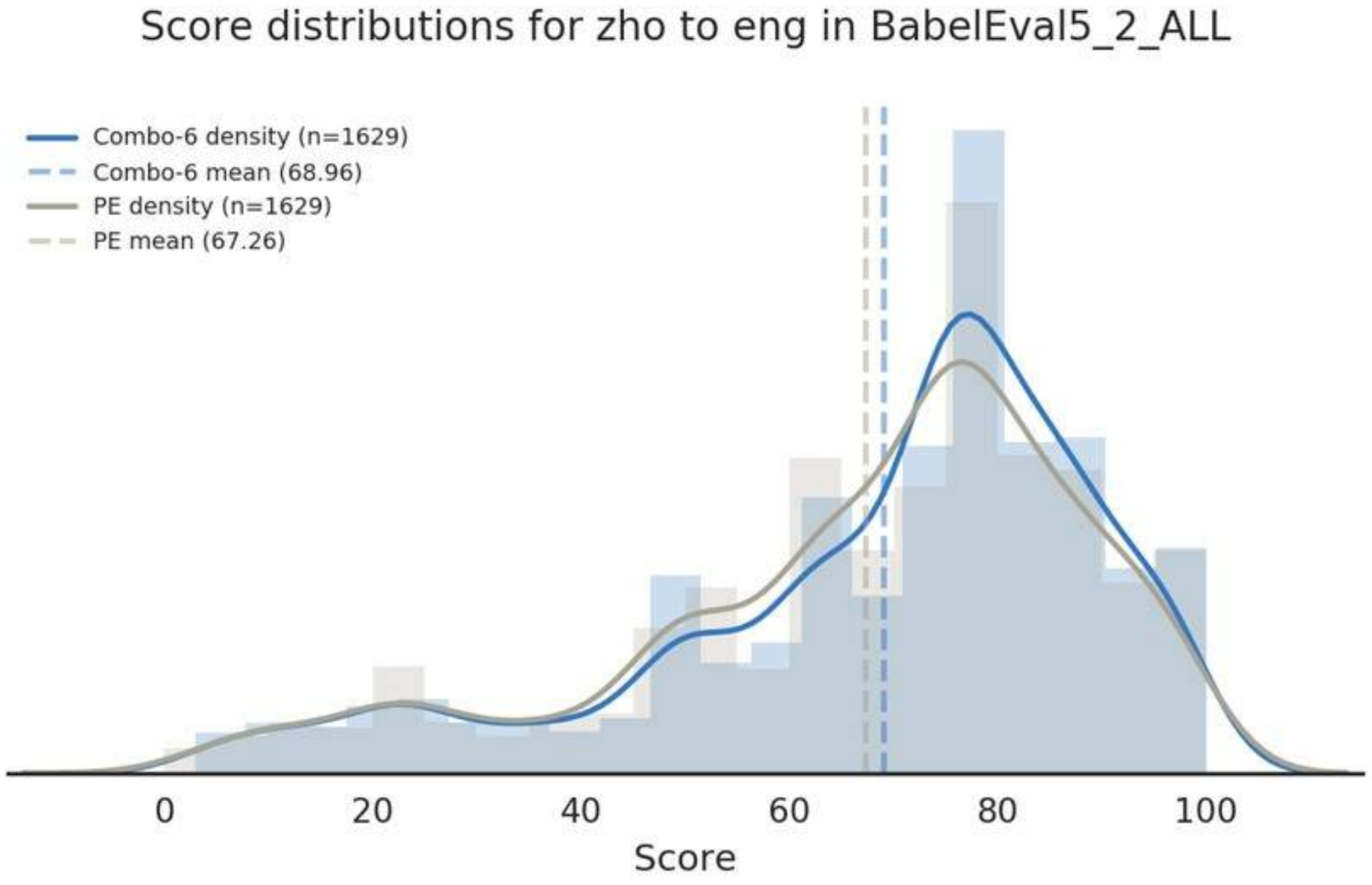
Combo-6 vs Sogou



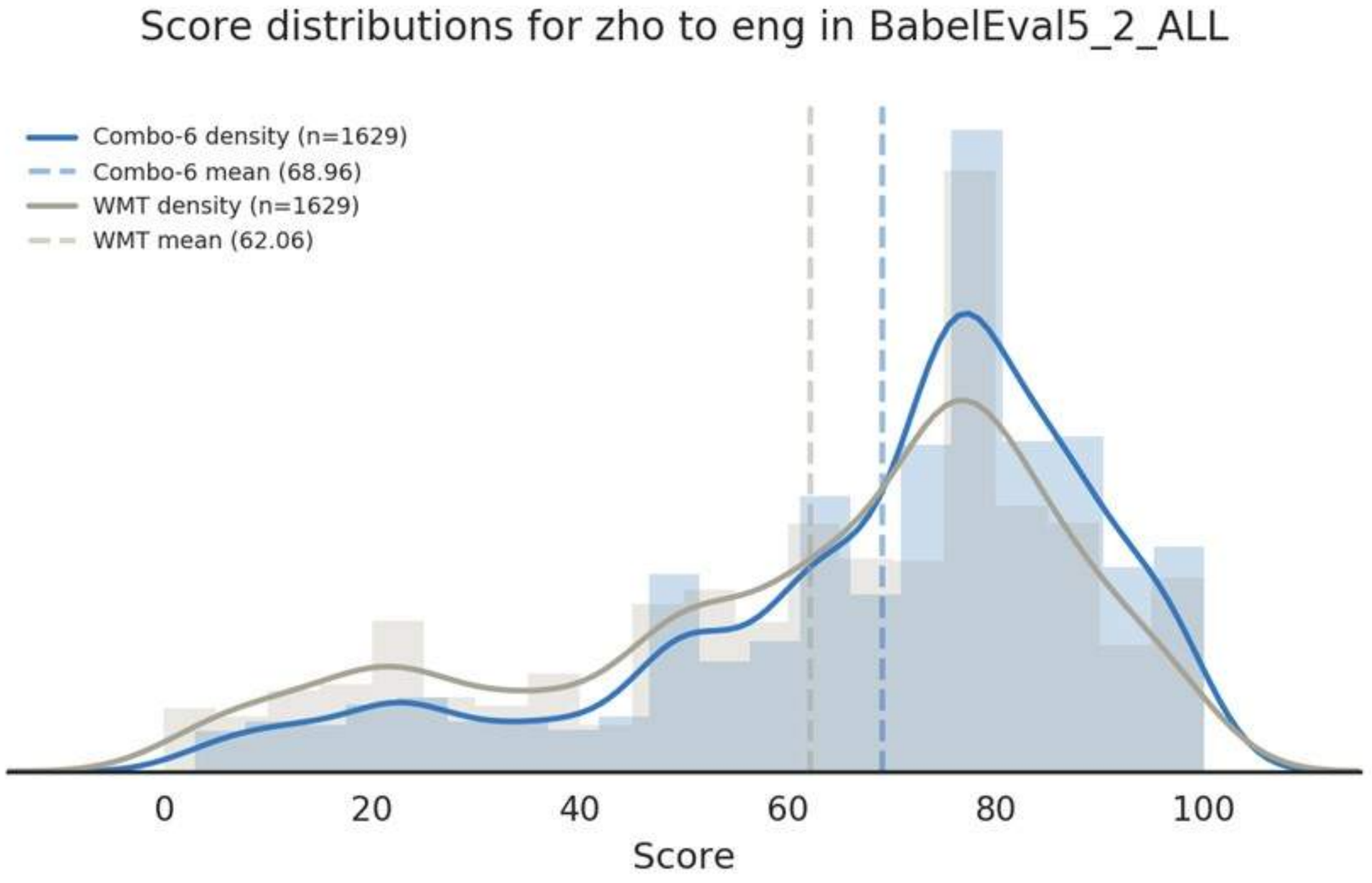
Combo-6 vs WMT



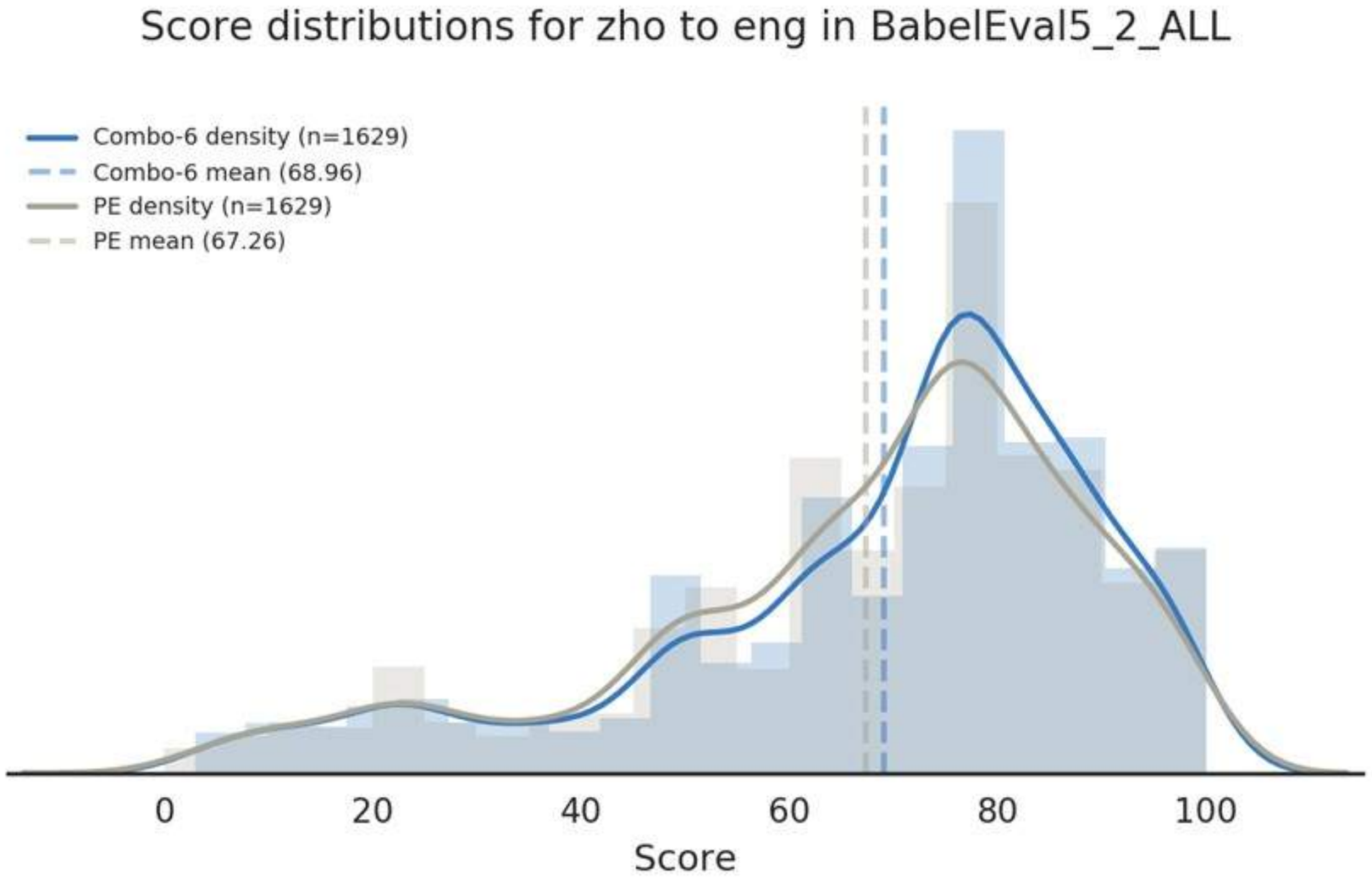
Combo-6 vs PE


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

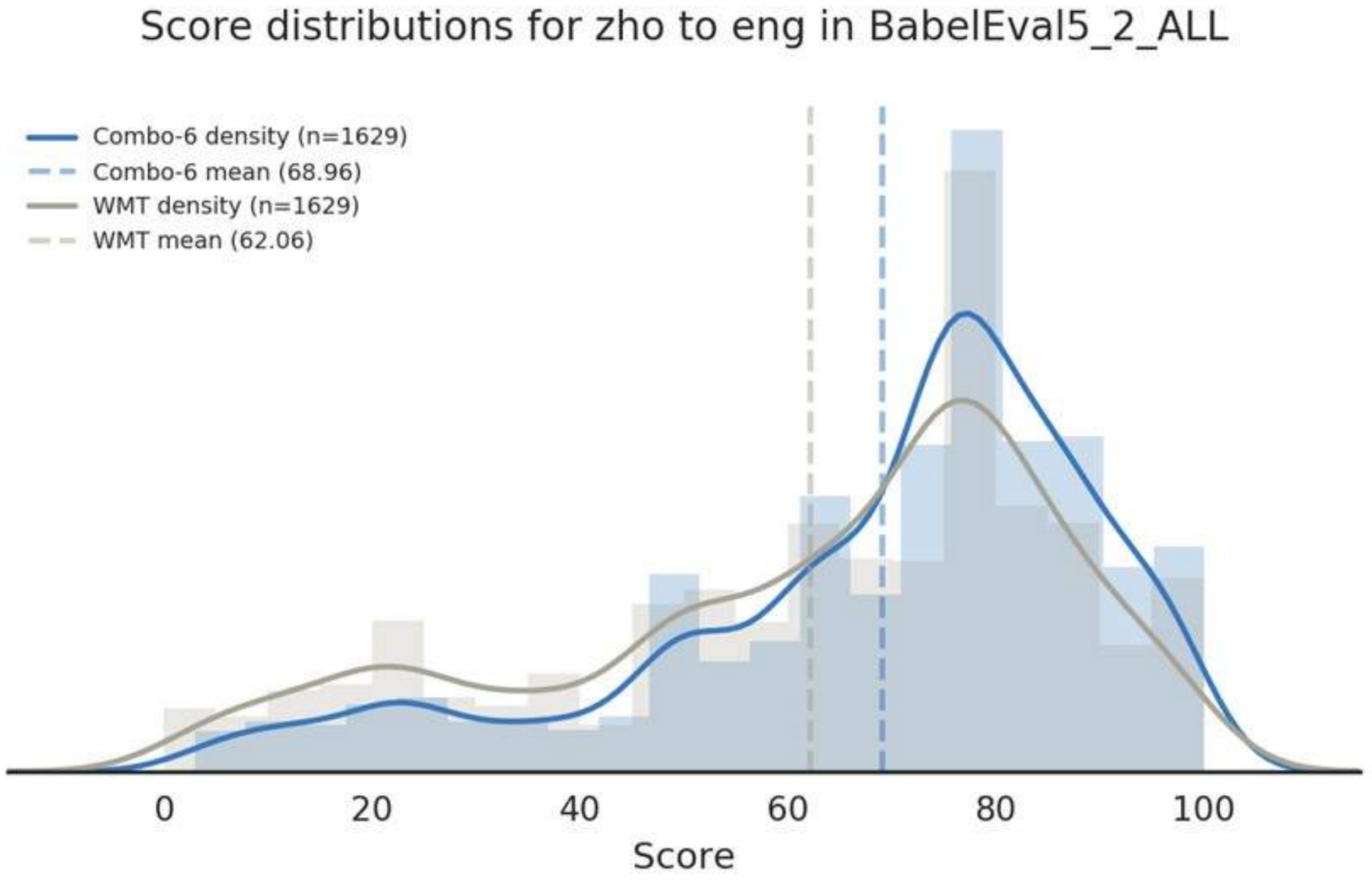
Combo-6 vs WMT



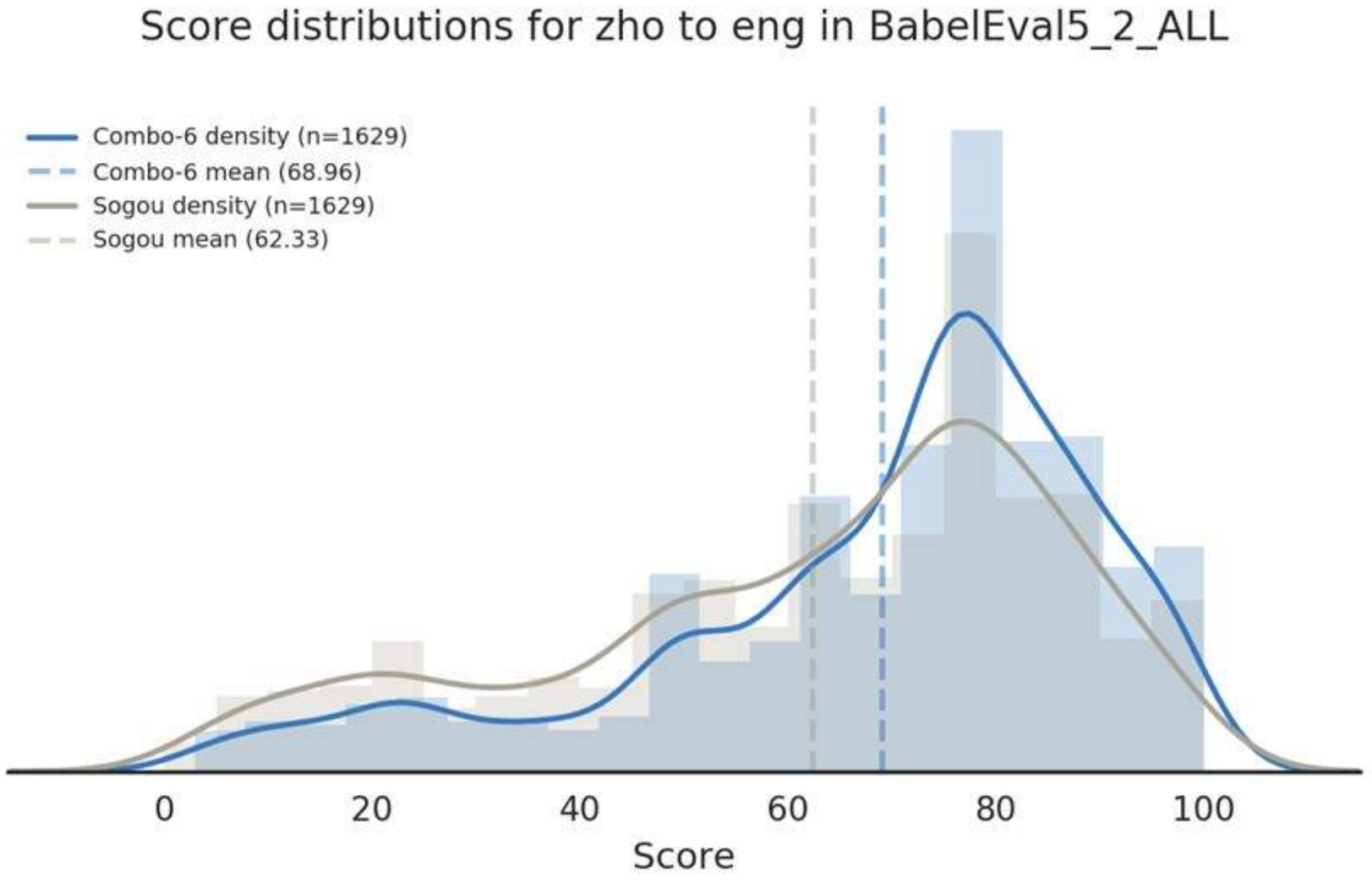
Combo-6 vs PE


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

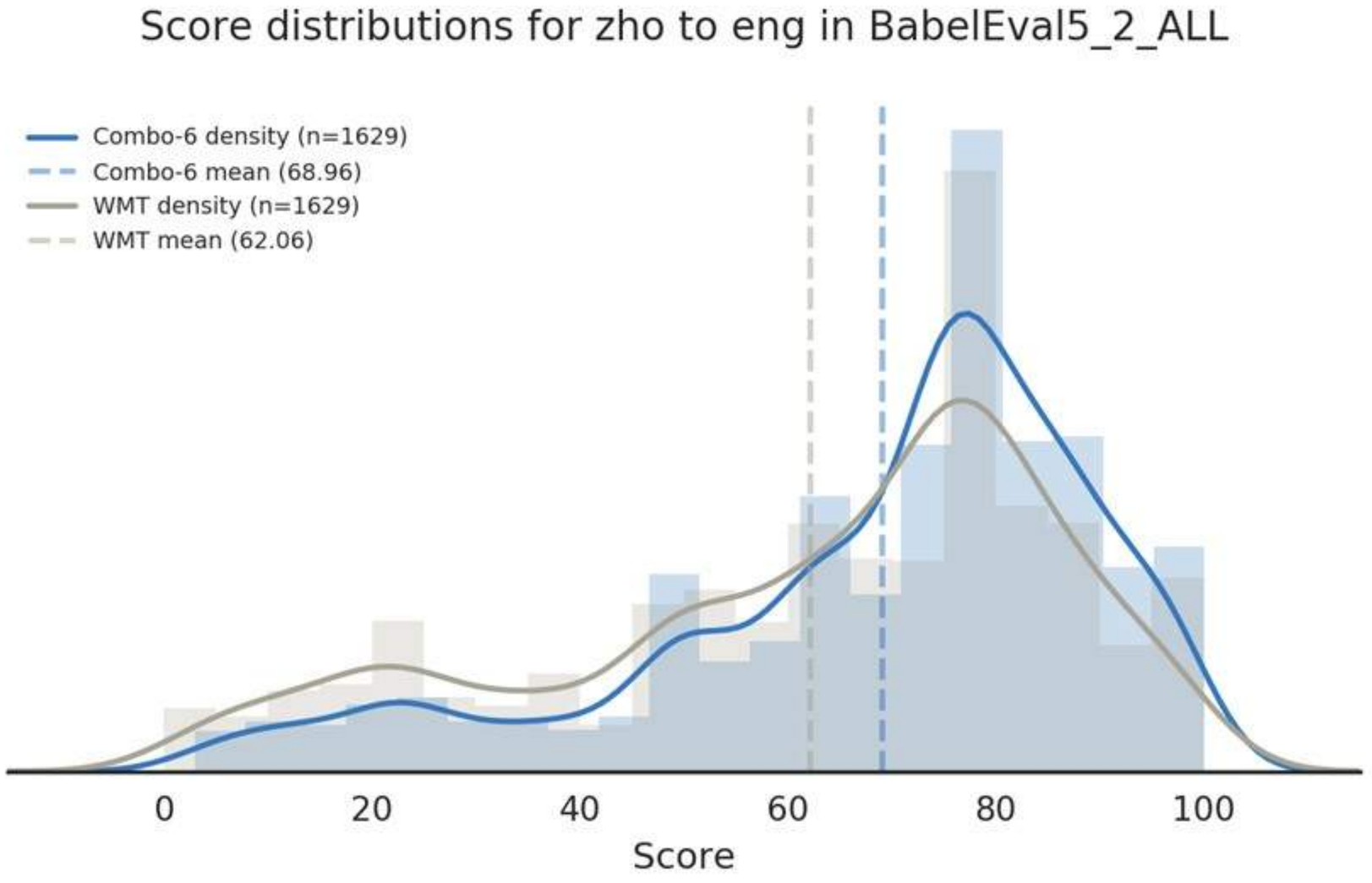
Combo-6 vs WMT



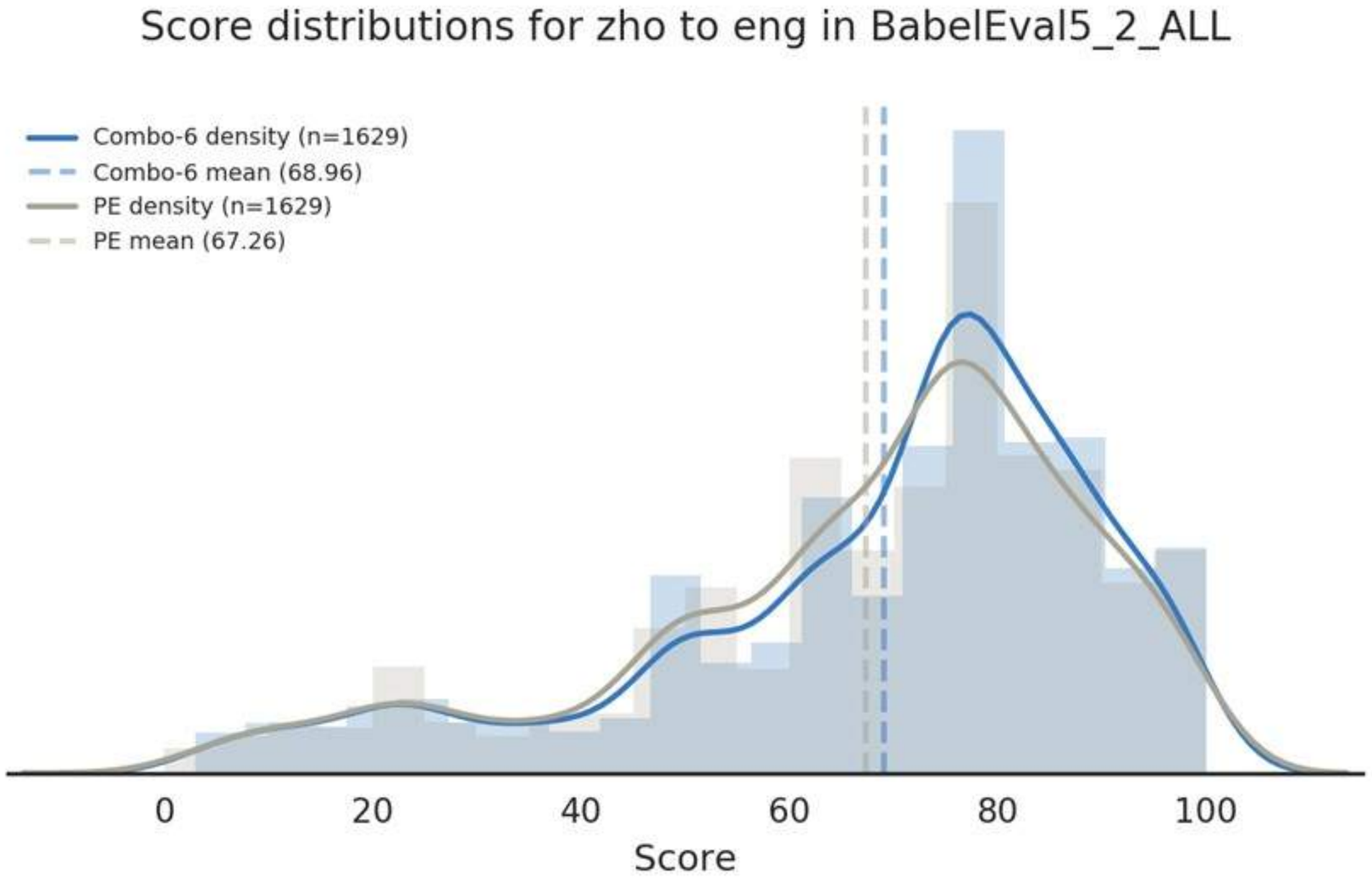
Combo-6 vs Sogou



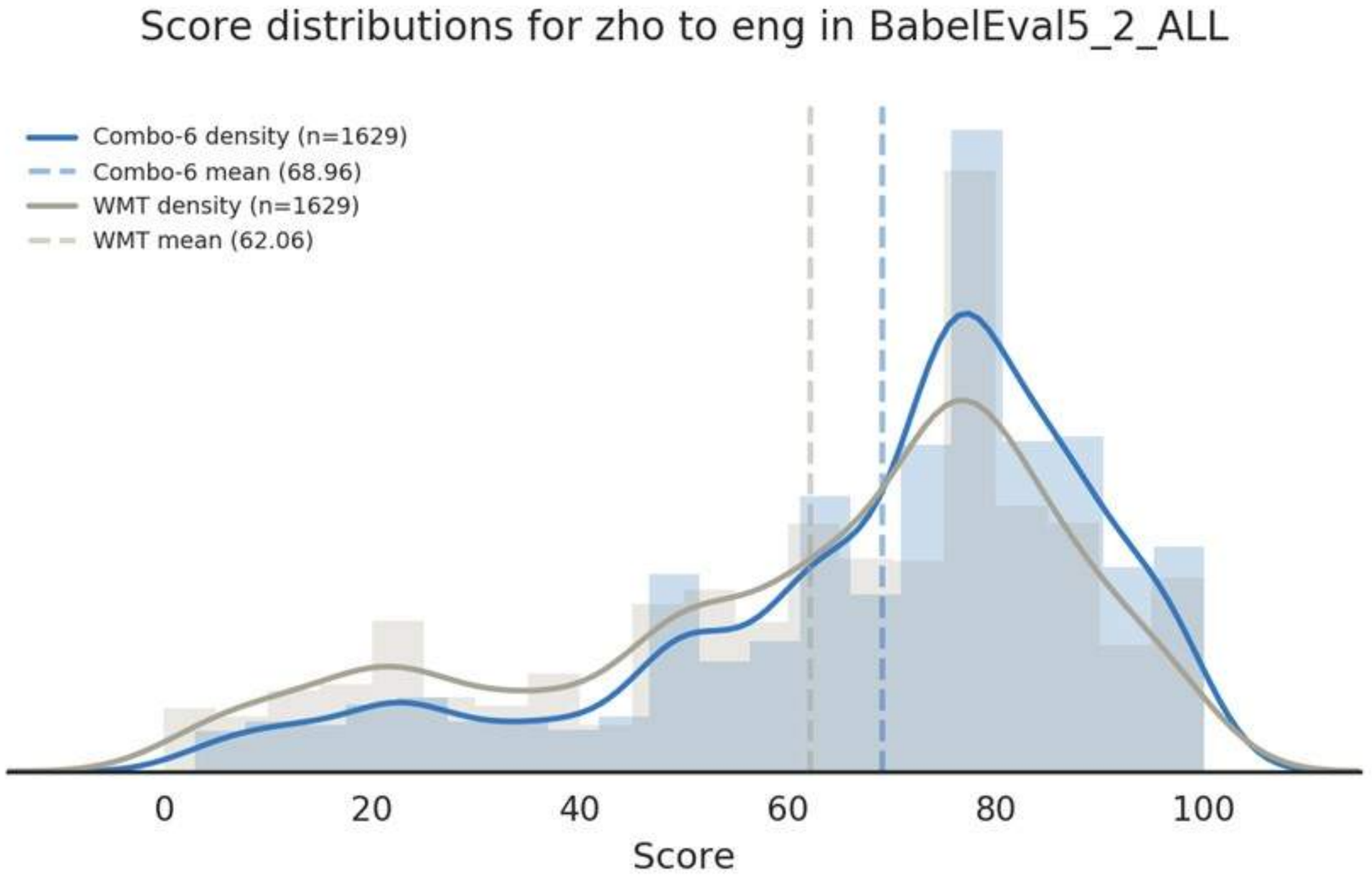
Combo-6 vs WMT



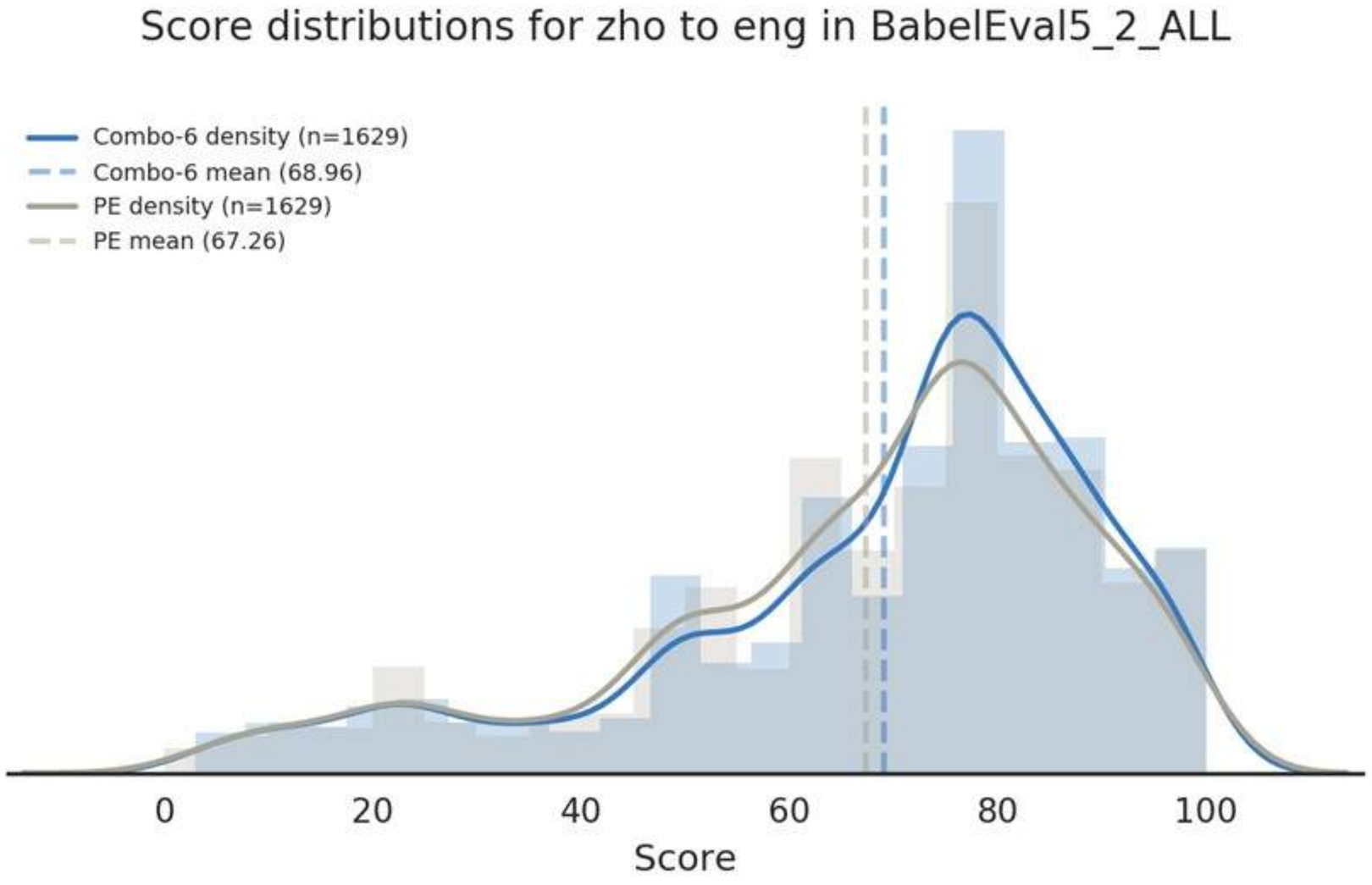
Combo-6 vs PE


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

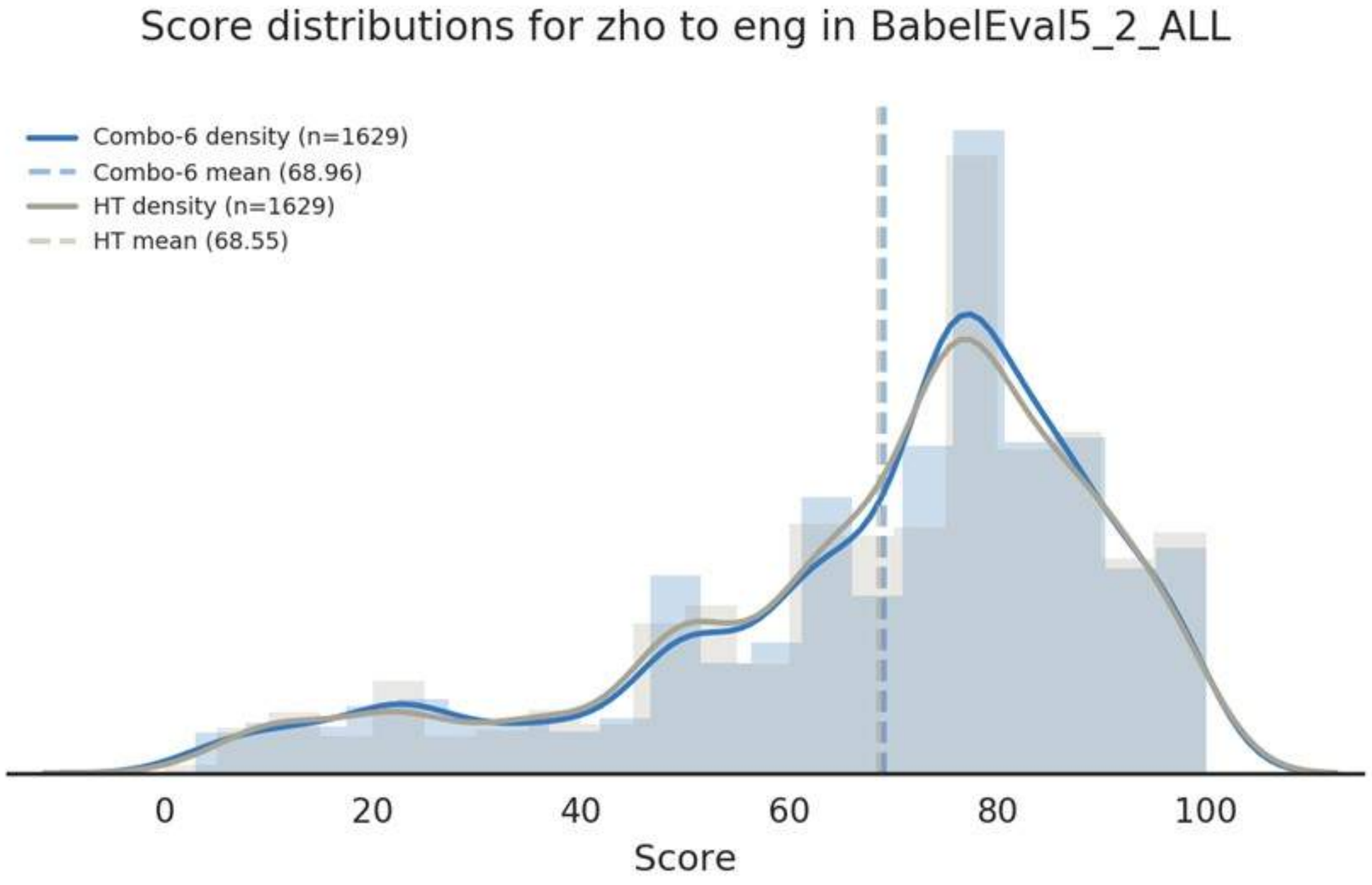
Combo-6 vs WMT



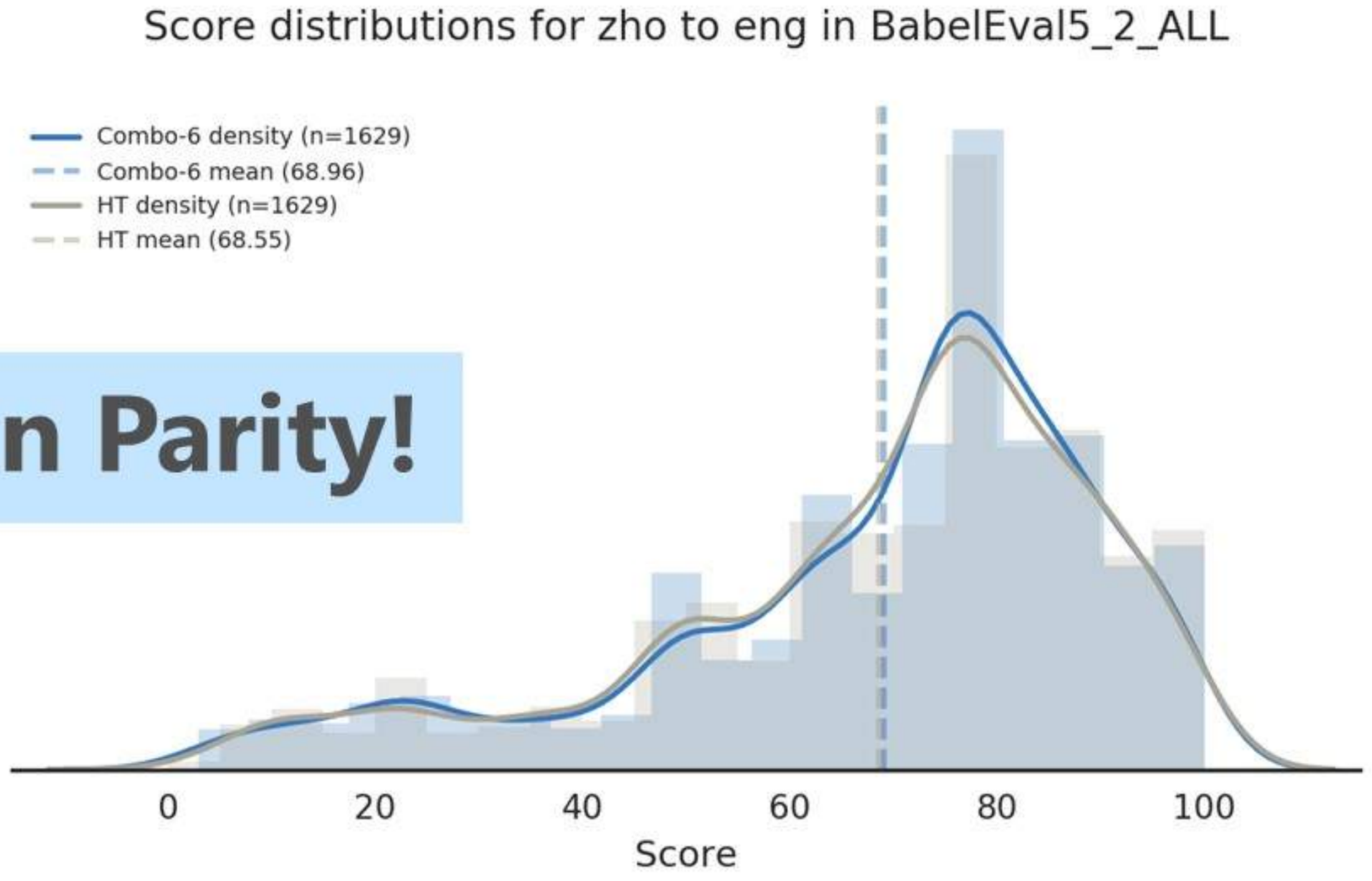
Combo-6 vs PE


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

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