Are we experiencing the Golden Age of Automatic Post-Editing?

Marcin Junczys-Dowmunt
Microsoft AI and Research

Translation Quality Estimation and Automatic Post-Editing
AMTA 2018

Why automatic post-editing?

Why automatic post-editing?

Can't we just retrain the original system?

Why automatic post-editing?

Can't we just retrain the original system?

Not always:

- black-box scenario
- specialized system make better use of PE data (?)
- synergy effects (RB-MT + SMT, SMT + NMT)

Popular metrics:
TER (Translation Error Rate) and BLEU

Historic APE systems:

Simard et. al (2007). Statistical Phrase-based Post-editing. NAACL.

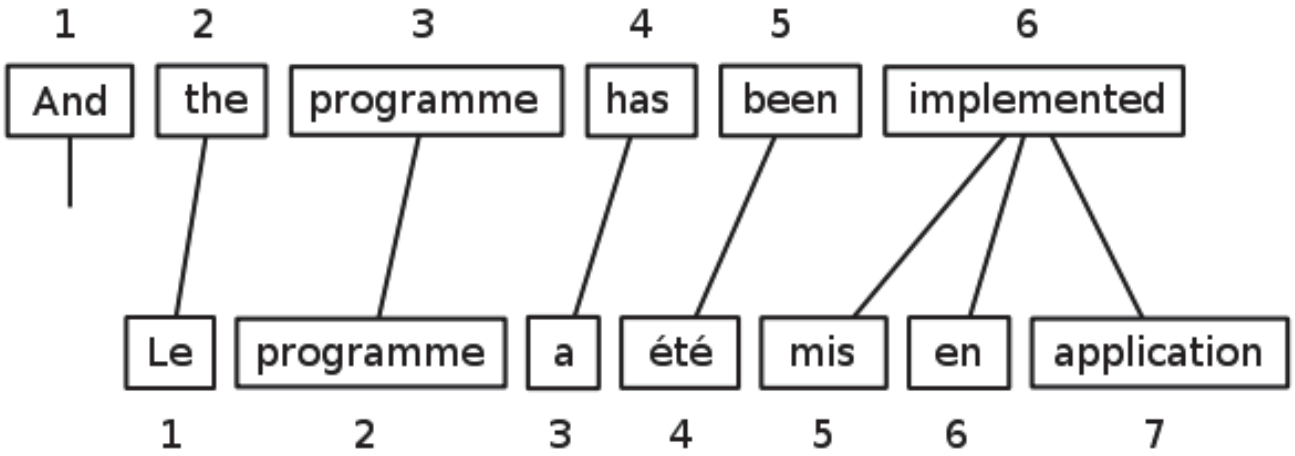
- Automatic Post-editing of a rule-based system with a phrase-based SMT system;
- About 30,000 paragraphs of triples per language pair (En-Fr/Fr-En);
- Train PB-SMT system on RB-MT output and PE data;
- Chain systems together;
- Impressive gains over the baselines.

Historic APE systems:

Bechara et. al (2011). Statistical Post-Editing for a Statistical MT System. MT-Summit.

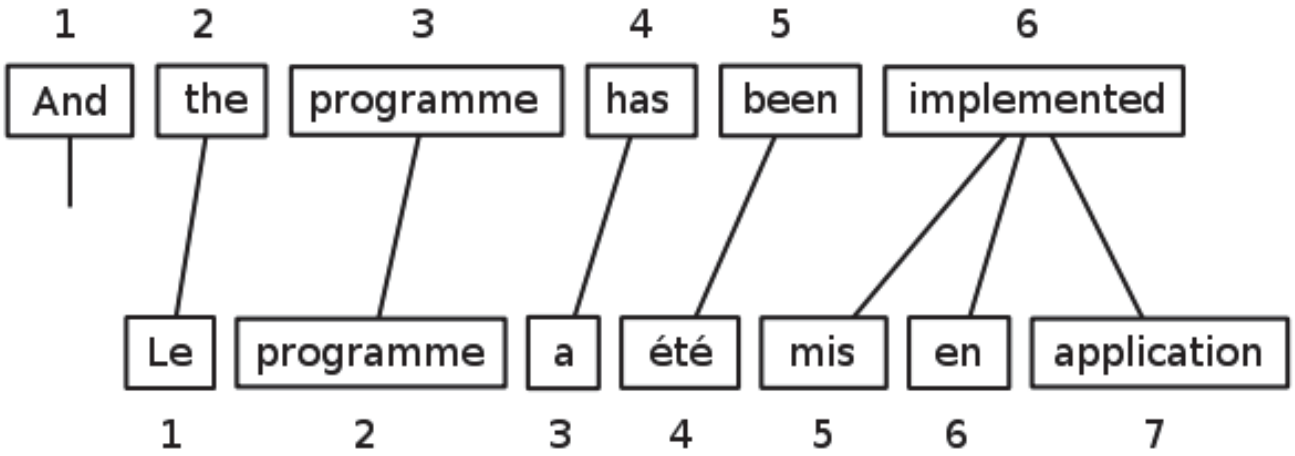
- Automatic Post-editing of a phrase-based SMT with another phrase-based SMT system.
- Barely any gains over the baselines.
- But interesting idea: Contextual Statistical APE

Contextual Statistical APE



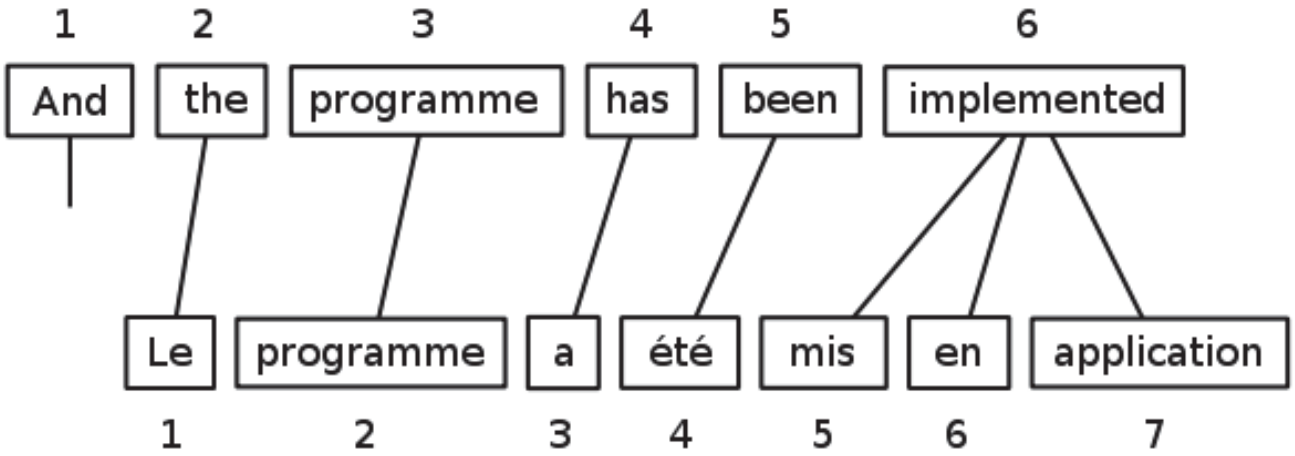
le#the

Contextual Statistical APE



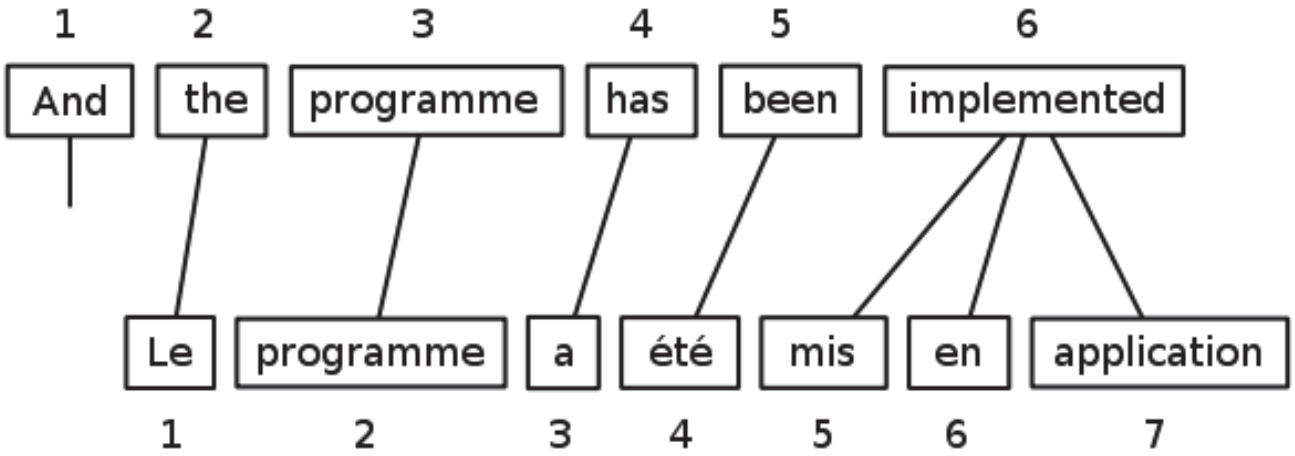
le#the programme#programme

Contextual Statistical APE



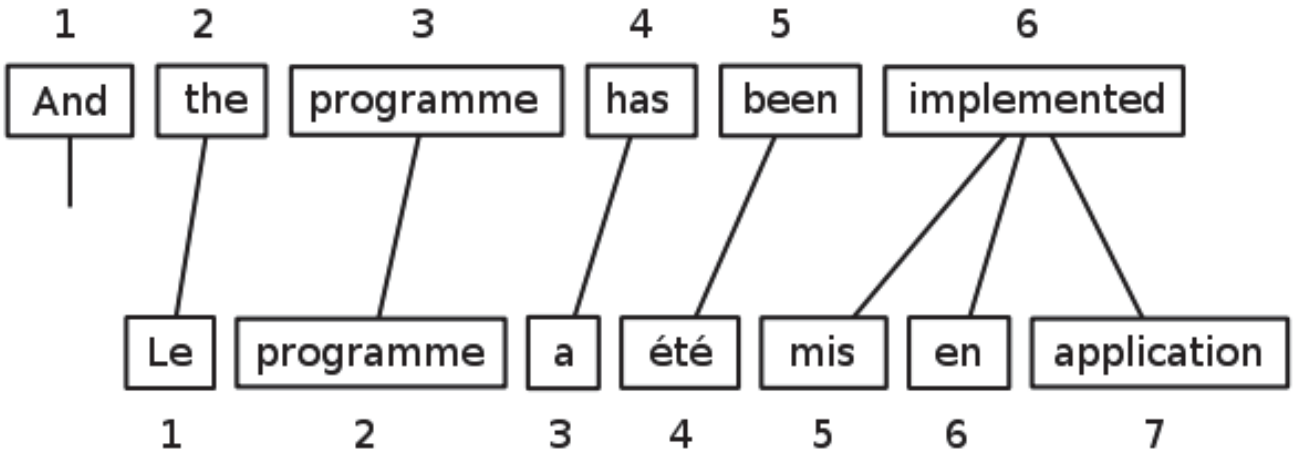
le#the programme#programme a#has

Contextual Statistical APE



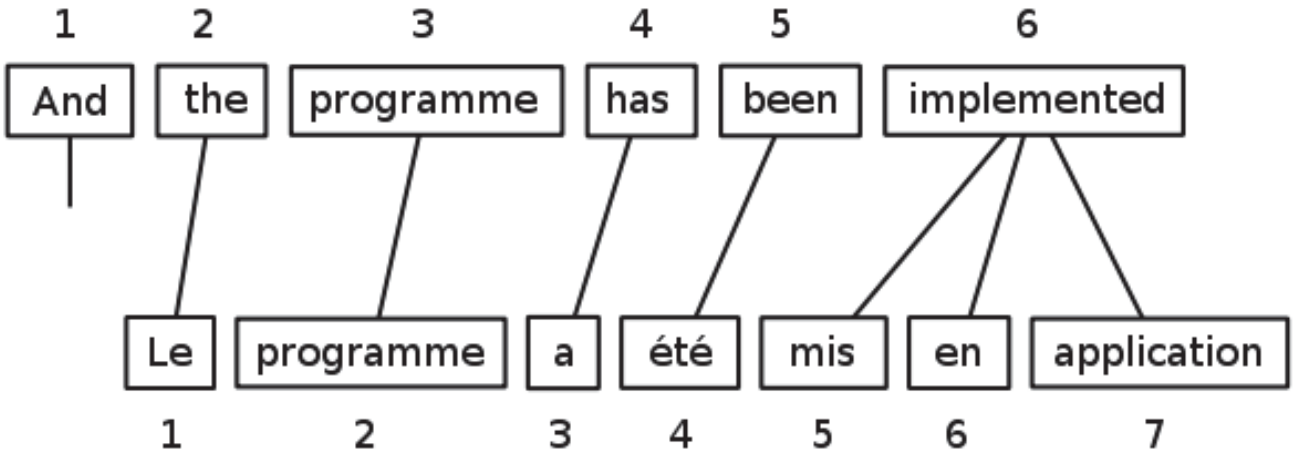
le#the programme#programme a#has été#been

Contextual Statistical APE



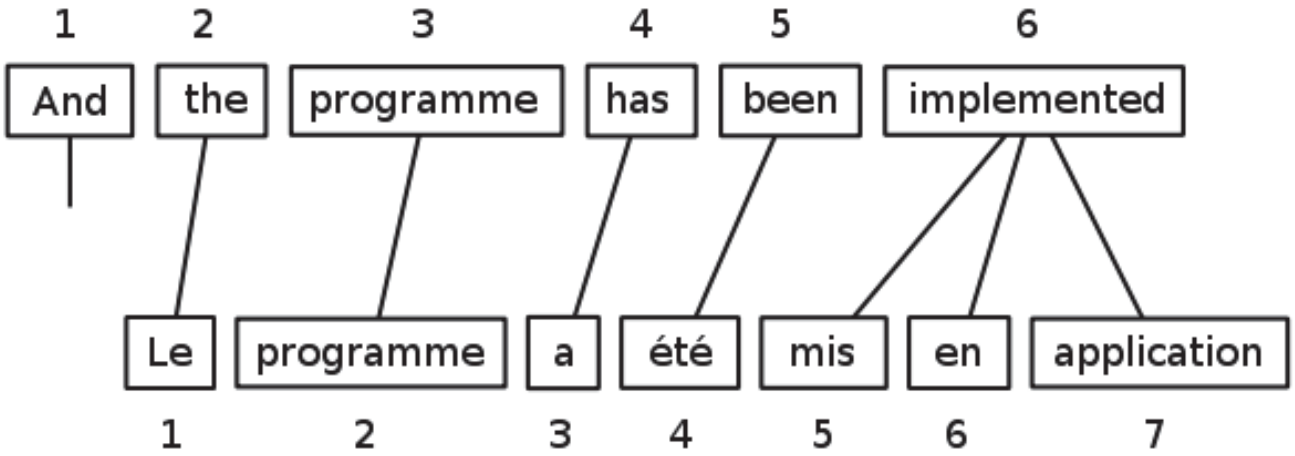
le#the programme#programme a#has été#been mis#implemented

Contextual Statistical APE



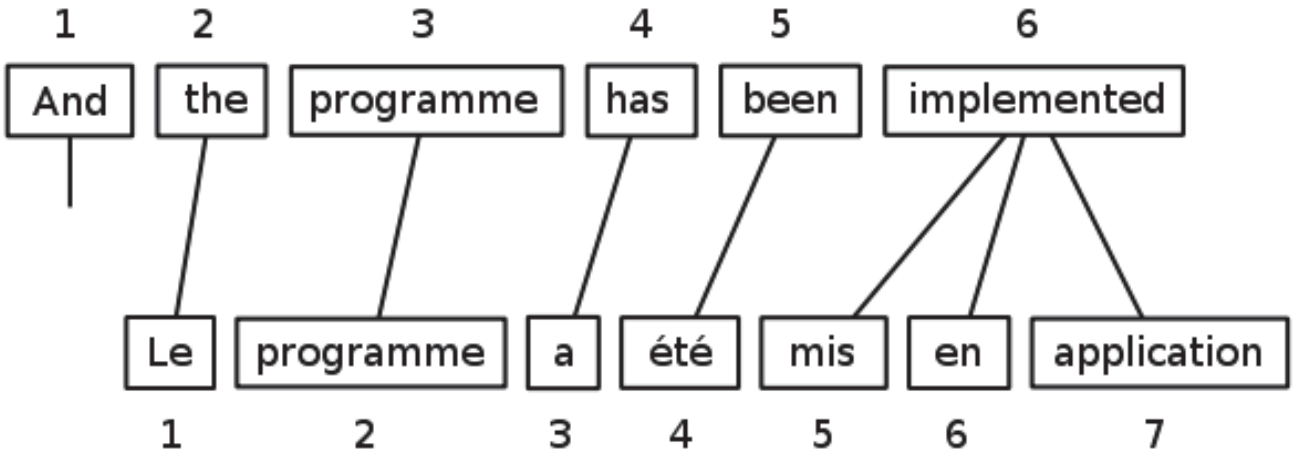
le#the programme#programme a#has été#been mis#implemented en#implemented

Contextual Statistical APE



le#the programme#programme a#has été#been mis#implemented en#implemented application#implemented

Contextual Statistical APE



le#the programme#programme a#has été#been mis#implemented en#implemented application#implemented

Problems?

WMT 2015 Shared Task on Automatic post-editing (The Stone Age of Automatic post-editing)

ID	Avg. TER
Baseline	22.91
FBK Primary	23.23
LIMSI Primary	23.33
USAAR-SAPE	23.43
LIMSI Contrastive	23.57
Abu-MaTran Primary	23.64
FBK Contrastive	23.65
(Simard et al., 2007)	23.84
Abu-MaTran Contrastive	24.72

WMT 2015 Shared Task on Automatic post-editing (The Stone Age of Automatic post-editing)

ID	Avg. TER
Baseline	22.91
FBK Primary	23.23
LIMSI Primary	23.33
USAAR-SAPE	23.43
LIMSI Contrastive	23.57
Abu-MaTran Primary	23.64
FBK Contrastive	23.65
(Simard et al., 2007)	23.84
Abu-MaTran Contrastive	24.72

WMT2016-best	23.29
WMT2017-best	??

WMT 2016 Shared Task on Automatic post-editing

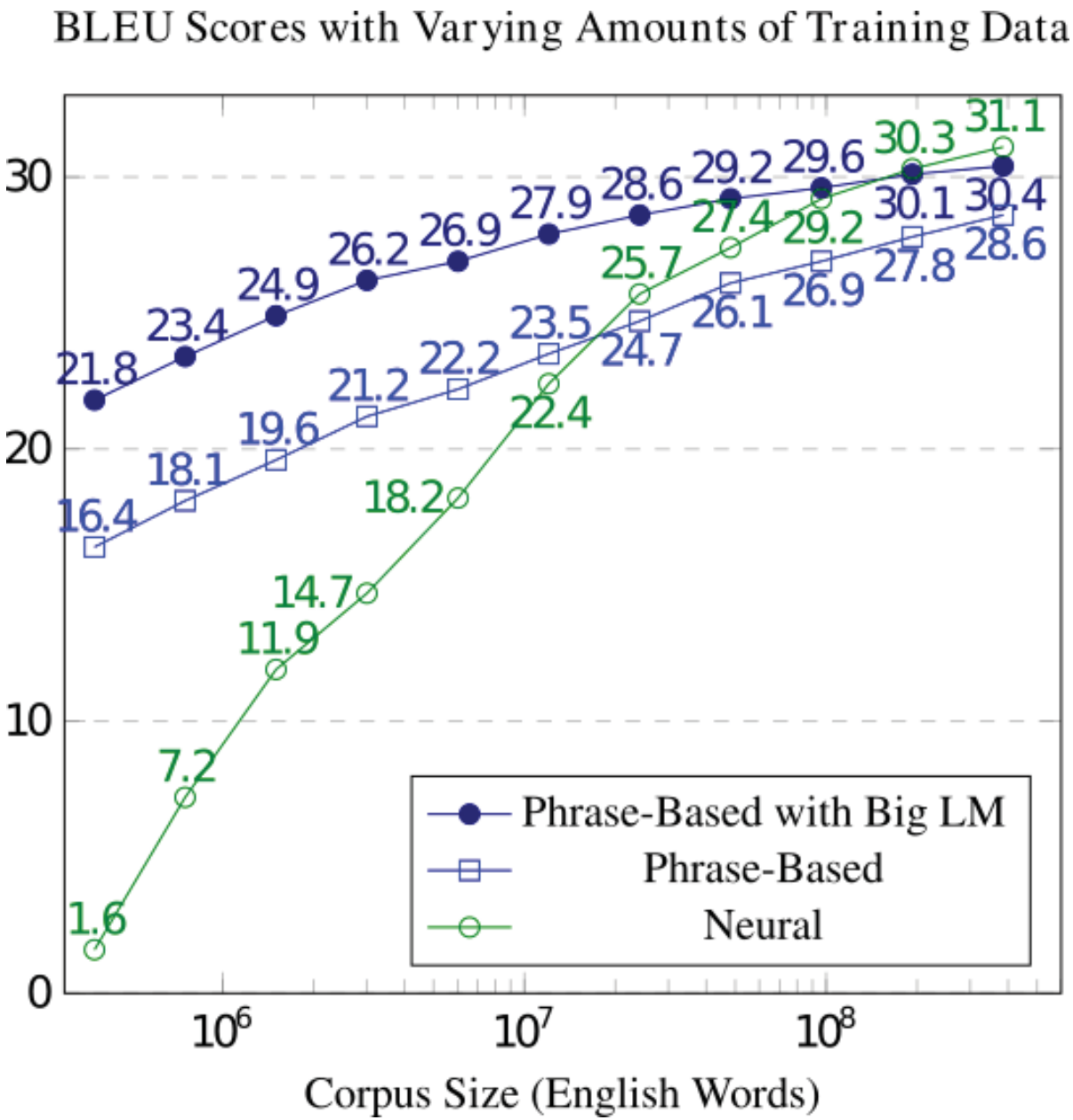
Create an APE system that returns automatic post-edition of an English-German black-box MT system. 10,000 training triplets of the following form were provided:

SRC *These files are encoded as UTF-8 or ASCII , which is a subset of UTF-8 .*

MT *Diese Dateien werden als UTF-8 oder ASCII , bei der es sich um eine Untergruppe von UTF-8 kodiert .*

PE *Diese Dateien werden als UTF-8 oder ASCII , eine Teilmenge von UTF-8 , kodiert .*

Problem: very little publicly available PE data



Source: Koehn and Knowles (2017). Six Challenges for Neural Machine Translation. 1st Neural Machine Translation Workshop. Vancouver.

Solution: create your own PE data using:

- Official APE training and develsopment data sets.
- EN-DE bilingual data from the WMT-16 shared tasks on IT and news translation.
- German monolingual Common Crawl (CC) corpus.

Round-trip translation

gibt die Prozesskennung des aktuellen Prozesses zurück . (= **PE**)

DE-EN ↓ Moses

the process ID of the current process . (= **SRC**)

EN-DE ↓ Moses

die Prozess-ID des aktuellen Prozesses . (= **MT**)

Selecting in-domain data

- Cross-entropy filtering of German CC corpus based on in-domain post-editing and IT-domain data.
- We keep 10M sentences with the best cross-entropy scores.

Filtering for TER statistics:

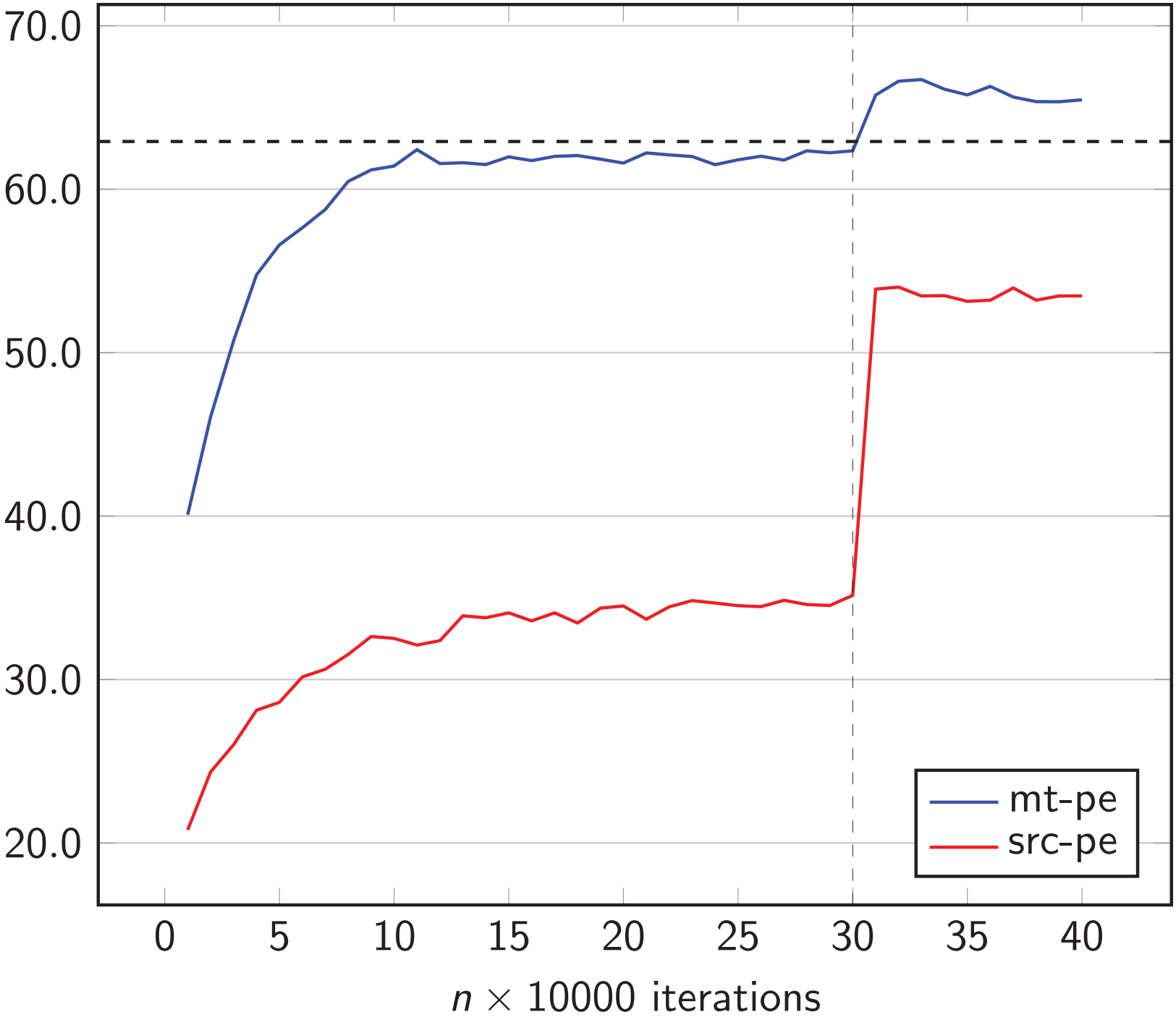
Data set	Sent.	NumWd	WdSh	NumEr	TER
training set	12K	17.89	0.72	4.69	26.22
development set	1K	19.76	0.71	4.90	24.81
round-trip.full	9,960K	13.50	0.58	5,72	42.02
round-trip.n10	4,335K	15.86	0.66	5.93	36.63
round-trip.n1	531K	20.92	0.55	5.20	25.28

Experiments with neural models

- Attentional encoded-decoder models trained with Nematus: `https://github.com/rsennrich/nematus`
- C++/CUDA AmuNMT decoder: `https://github.com/emjotde/amunmt`

MT-PE and SRC-PE systems

- Trained on round-trip.n10 data (4M triplets).
- Fine-tuned on round-trip.n1 and 20x oversampled official training data (700K triplets).


70.0
60.0
50.0
40.0
30.0
20.0
0
5
10
15
20
25
30
35
40
$n \times 10000$ iterations
mt-pe
src-pe

Log-linear combination

- Log-linear combination of two models with different input languages.
- Weights determined by MERT for two models: ca. 0.8 for mt-pe and 0.2 for src-pe model.
- Post-Editing Penalty (PEP) to control the faithfulness of the APE results.

Progress on the dev set

System	TER	BLEU
Baseline (mt)	25.14	62.92
mt→pe	23.37	66.71
mt→pe×4	23.23	66.88
src→pe	32.31	53.89
src→pe×4	31.42	55.41
mt→pe×4 / src→pe×4	22.38	68.07
mt→pe×4 / src→pe×4 / pep	**21.46**	**68.94**

Automatic evaluation on unseen test set

- AMU (primary) = mt→pe×4 / src→pe×4 / pe
- AMU (contrastive) = mt→pe×4

System	TER	BLEU
AMU (primary)	**21.52**	**67.65**
AMU (contrastive)	**23.06**	**66.09**
FBK	23.92	64.75
USAAR	24.14	64.10
CUNI	24.31	63.32
Baseline (Moses)	24.64	63.47
Baseline (mt)	24.76	62.11
DCU	26.79	58.60
JUSAAR	26.92	59.44

Results of human evaluation

#	Score	Range	System
1	**1.967**	1	**AMU (primary)**
2	0.033	2	FBK
3	-0.108	3-4	CUNI
	-0.191	3-5	USSAR
	-0.211	3-5	Baseline (mt)
4	-0.712	6-7	JUSAAR
	-0.778	6-7	DCU

Table: With post-edited sentence shown as reference

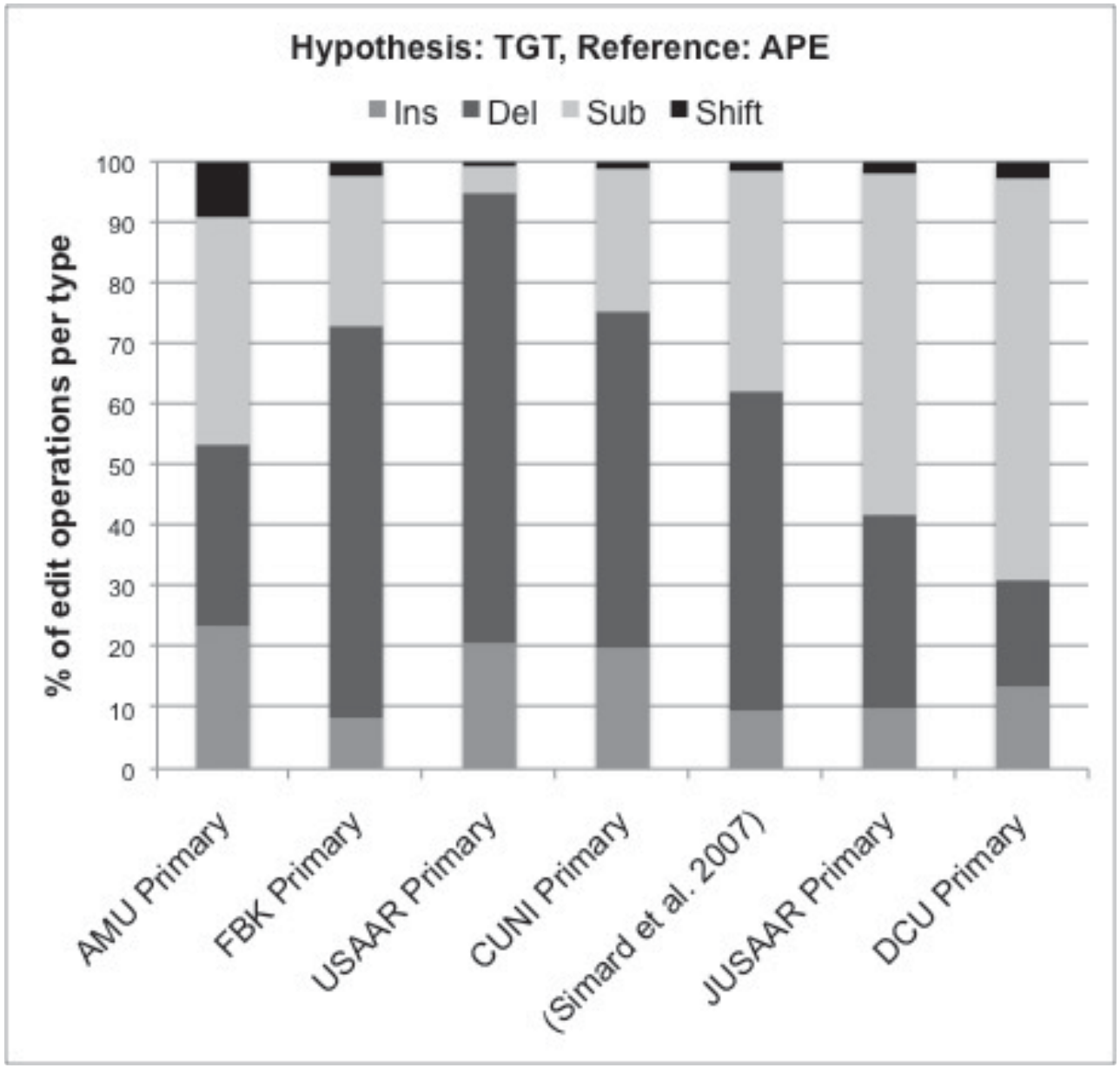
Source: WMT2016 overview paper.

Results of human evaluation

#	Score	Range	System
1	2.058	1	Human
2	**0.867**	2	**AMU (primary)**
3	-0.213	3-4	CUNI
	-0.348	3-6	FBK
	-0.374	3-6	USSAR
	-0.499	5-7	Baseline (mt)
	-0.675	6-8	JUSAAR
	-0.816	7-8	DCU

Table: With post-edited sentence included as system

Source: WMT2016 overview paper.



Source: WMT 2016 overview paper

Some conclusions

- One of the first successful applications of NMT models to APE
- Artificial APE triplets allow training of NMT models with little original training data and help against overfitting.
- Positive effects of log-linear combinations of NMT models with multiple input languages.
- Tuning with MERT to assign model component weights

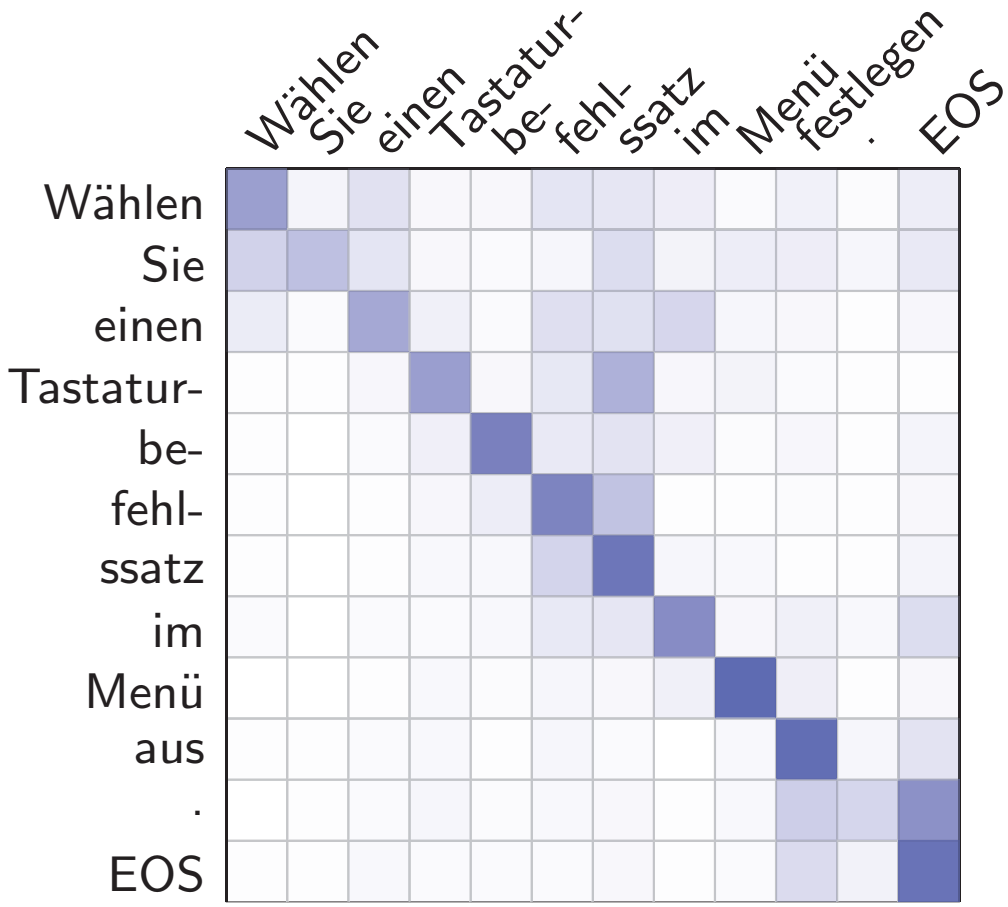
WMT 2017 Shared Task on Automatic post-editing

- The same setting;
- Additional 12,500 sentences of PE data;
- Still no post-editing of NMT system

Our submission to the WMT 2017 Shared Task on Automatic Post-editing

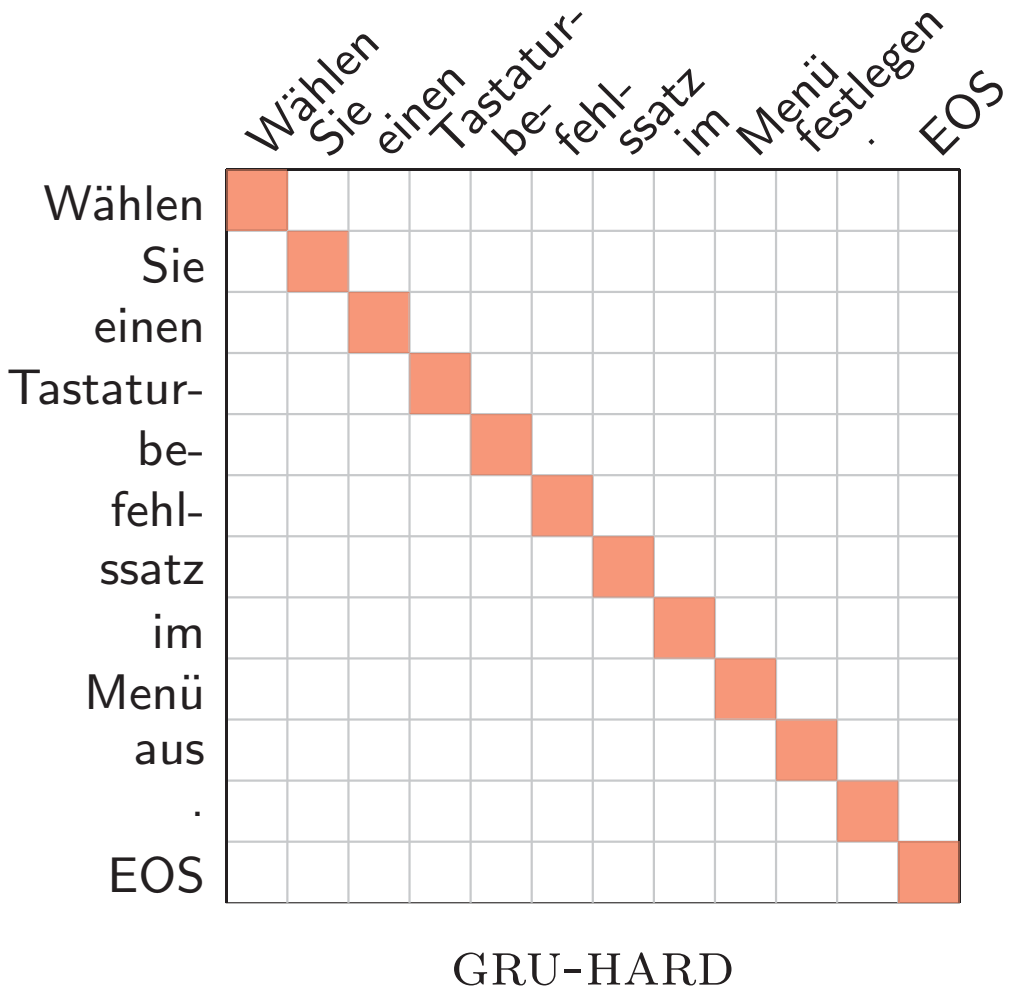
- We explore the interaction of hard-attention and multi-encoder models.
- All models trained and available in Marian (http://marian-nmt.github.io)
- We use the same data as last year.
- This time proper regularization and no need for fine-tuning.

Soft vs. hard monotonic attention



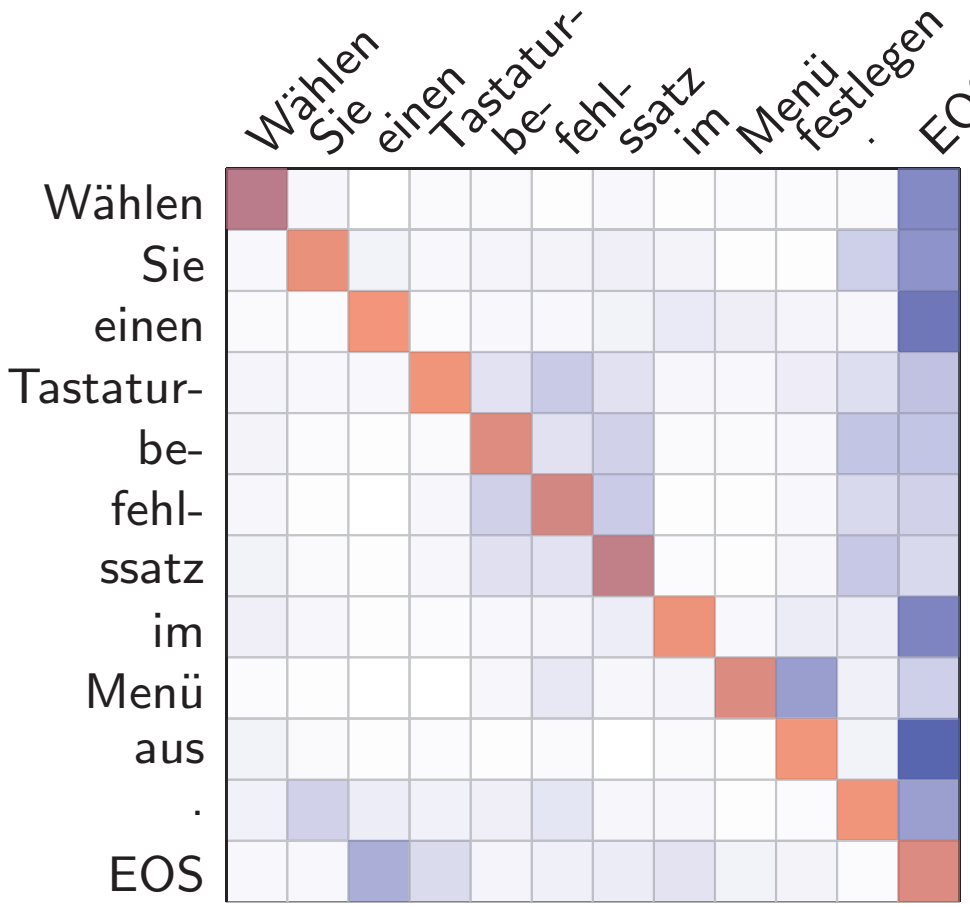
CGRU

Soft vs. hard monotonic attention



GRU-HARD

Soft vs. hard monotonic attention



CGRU-HARD

Reminder: Gated Recurrent Unit

$$
\begin{aligned}
\mathrm{GRU}\left(\mathbf{s}, \mathbf{x}\right) =& (1 - \mathbf{z}) \odot \underline{\mathbf{s}} + \mathbf{z} \odot \mathbf{s}, \\
\underline{\mathbf{s}} =& \tanh\left(\mathbf{W}\mathbf{x} + \mathbf{r} \odot \mathbf{U}\mathbf{s}\right), \\
\mathbf{r} =& \sigma\left(\mathbf{W}_r\mathbf{x} + \mathbf{U}_r\mathbf{s}\right), \\
\mathbf{z} =& \sigma\left(\mathbf{W}_z\mathbf{x} + \mathbf{U}_z\mathbf{s}\right),
\end{aligned}
\tag{1}
$$

where $\mathbf{x}$ is the cell input; $\mathbf{s}$ is the previous recurrent state; $\mathbf{W}$, $\mathbf{U}$, $\mathbf{W}_r$, $\mathbf{U}_r$, $\mathbf{W}_z$, $\mathbf{U}_z$ are trained model parameters[1]; σ is the logistic sigmoid activation function.

[1] Biases have been omitted.

Conditional GRU (cgru)

$$C = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$$

$$\mathbf{s}_j = \text{cGRU}_{\text{att}}\left(\mathbf{s}_{j-1}, \mathbf{E}[y_{j-1}], \text{C}\right) \tag{2}$$

$$\begin{aligned}
\mathbf{s}j' &= \text{GRU}_1\left(\mathbf{s}_{j-1}, \mathbf{E}[y_{j-1}]\right) \\
\mathbf{c}_j &= \text{ATT}\left(\text{C}, \mathbf{s}'_j\right) \\
\mathbf{s}_j &= \text{GRU}_2\left(\mathbf{s}'_j, \mathbf{c}_j\right)
\end{aligned}$$

Hard monotonic attention (gru-hard)

- Aharoni and Goldberg (2016) introduce a simple model for monolingual morphological re-inflection with hard monotonic attention.
- The target word vocabulary V_y is extended with a special step symbol $\langle \text{STEP} \rangle$
- Whenever $\langle \text{STEP} \rangle$ is predicted as the output symbol, the hard attention is moved to the next encoder state.
- We calculate the hard attention indices as follows:

$$a_1 = 1,$$

$$a_j = \begin{cases} a_{j-1} + 1 & \text{if } y_{j-1} = \langle \text{STEP} \rangle \\ a_{j-1} & \text{otherwise.} \end{cases}$$

$$\mathbf{s}_j = \text{GRU}\left(\mathbf{s}_{j-1}, \left[\mathbf{E}[y_{j-1}]; \mathbf{h}_{a_j}\right]\right), \tag{3}$$

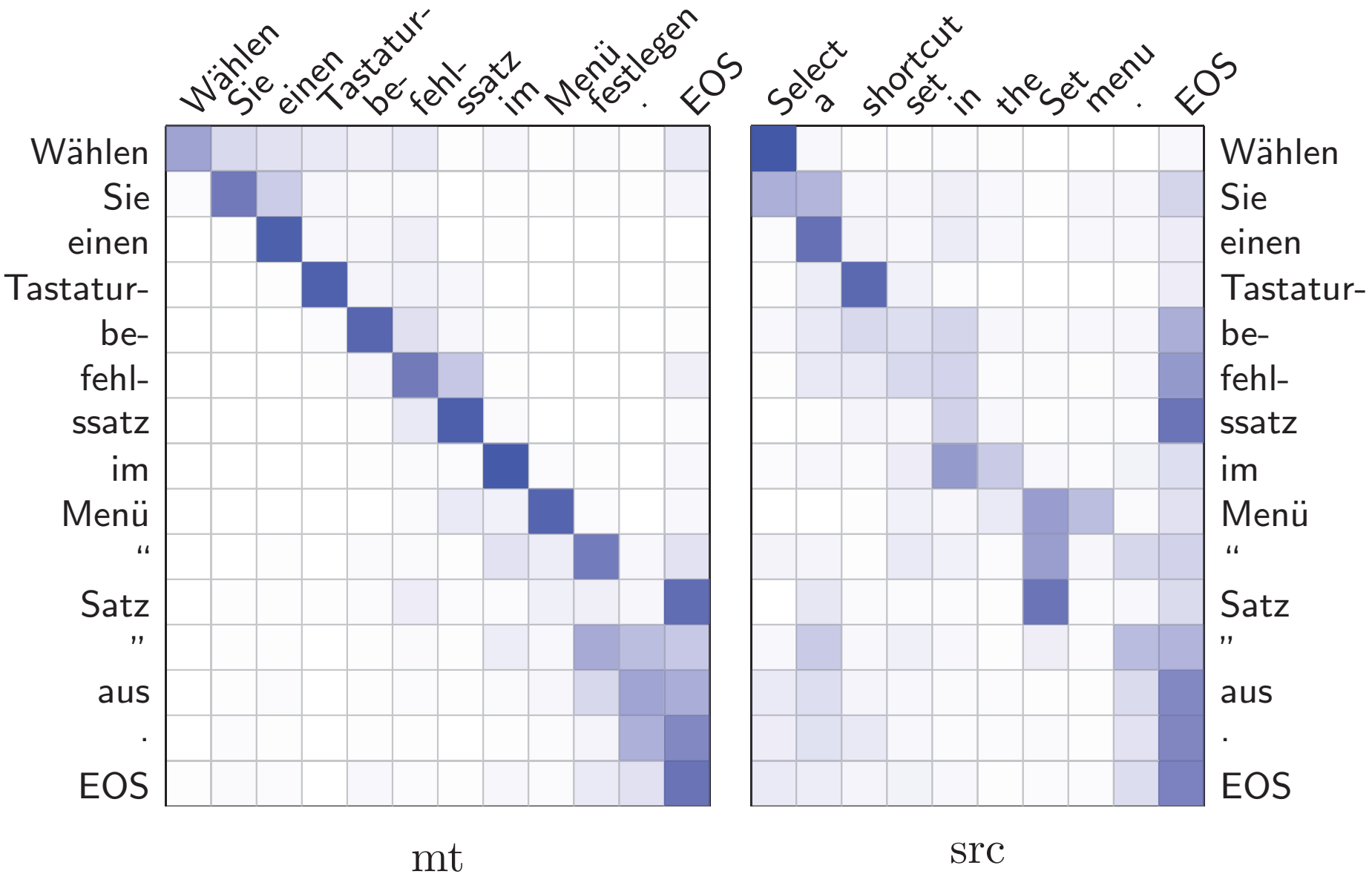
Mixing hard and soft attention (cgru-hard)

$$\mathbf{s}_j = \mathrm{cGRU}_{\mathrm{att}}\left(\mathbf{s}_{j-1}, \left[\mathbf{E}[y_{j-1}]; \mathbf{h}_{a_j}\right], \mathrm{C}\right) \tag{4}$$

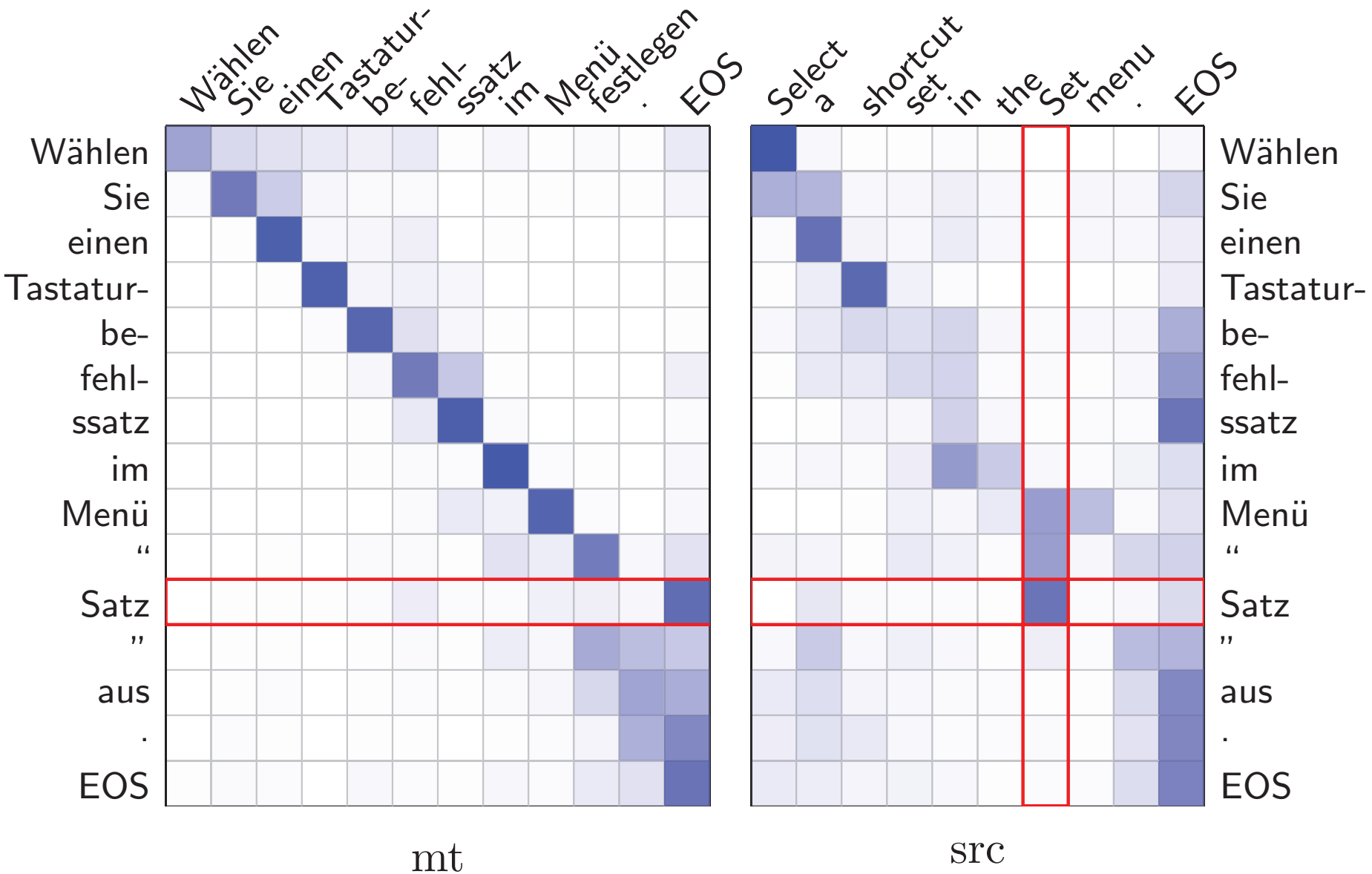
Example sentence and corrections

mt	Wählen Sie einen Tastaturbefehlssatz im Menü festlegen .
src	Select a shortcut set in the Set menu .
CGRU	Wählen Sie einen Tastaturbefehlssatz im Menü aus .
GRU-HARD	Wählen Sie einen Tastaturbefehlssatz im Menü aus .
CGRU-HARD	Wählen Sie einen Tastaturbefehlssatz im Menü aus .
M-CGRU	Wählen Sie einen Tastaturbefehlssatz im Menü " Satz " aus .
M-CGRU-HARD	Wählen Sie einen Tastaturbefehlssatz im Menü " Satz . "
pe	Wählen Sie einen Tastaturbefehlssatz im Menü " Satz . "

Dual attention



Dual attention



Dual soft attention (m-cgru)

$$C^{mt} = \{\mathbf{h}_1^{mt}, \ldots, \mathbf{h}_{T_{mt}}^{mt}\}$$
$$C^{src} = \{\mathbf{h}_1^{src}, \ldots, \mathbf{h}_{T_{src}}^{src}\}$$

$$\mathbf{s}_0 = \tanh\left(\mathbf{W}_{init}\left[\frac{\sum_{i=1}^{T_{mt}}\mathbf{h}_i^{mt}}{T_{mt}}; \frac{\sum_{i=1}^{T_{src}}\mathbf{h}_i^{src}}{T_{src}}\right]\right).$$

$$\mathbf{s}_j = \text{cGRU}_{\text{2-att}}\left(\mathbf{s}_{j-1}, \mathbf{E}[y_{j-1}], C^{mt}, C^{src}\right). \tag{5}$$

Dual soft attention (m-cgru)

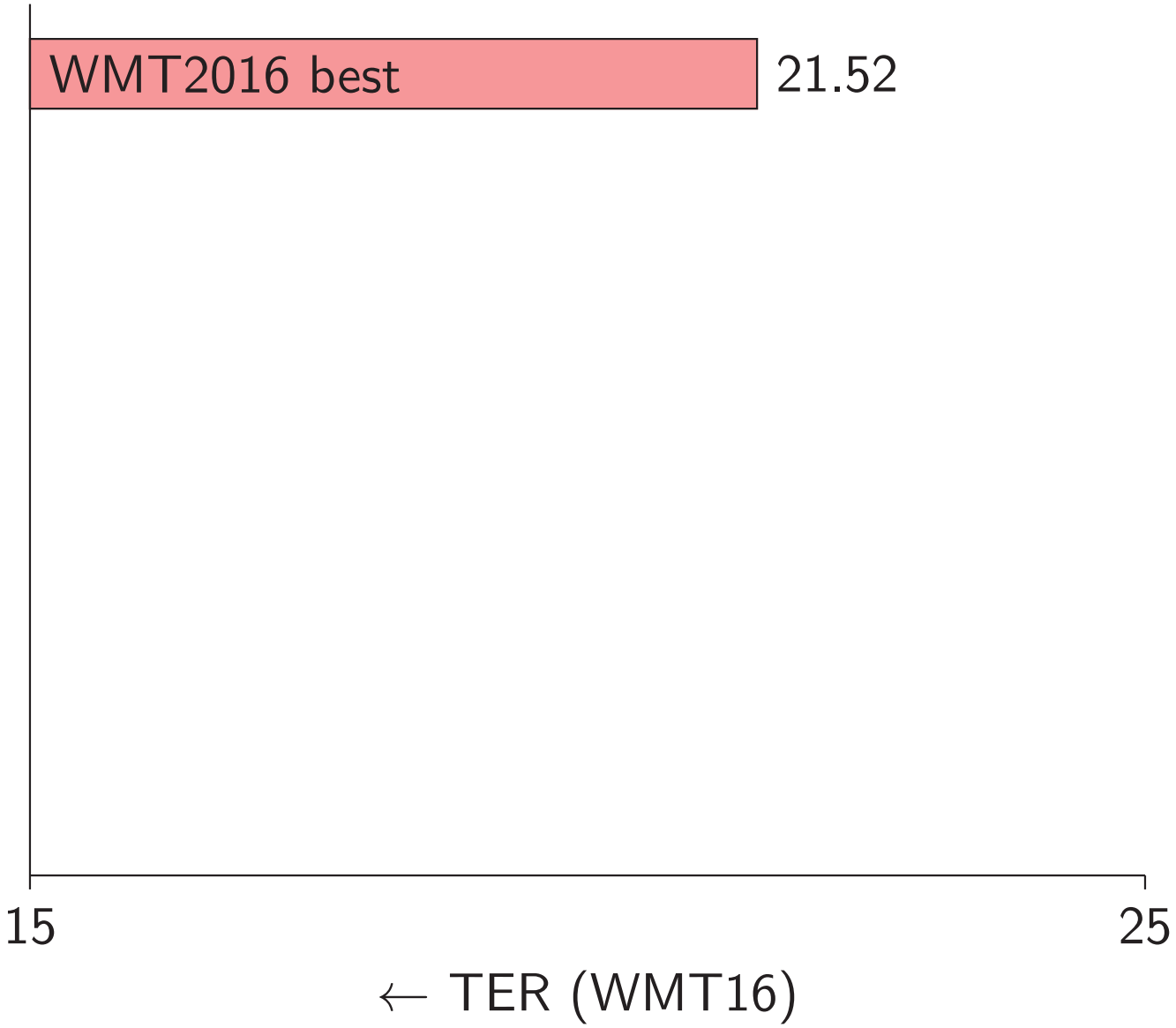
$$\mathbf{s}_j = \text{cGRU}_{\text{2-att}} \left(\mathbf{s}_{j-1}, \mathbf{E}[y_{j-1}], C^{mt}, C^{src} \right). \tag{6}$$

$$\begin{aligned}
\mathbf{s}'_j &= \text{GRU}_1 \left(\mathbf{s}_{j-1}, \mathbf{E}[y_{j-1}] \right), \\
\mathbf{c}^{mt}_j &= \text{ATT} \left(C^{mt}, \mathbf{s}'_j \right), \\
\mathbf{c}^{src}_j &= \text{ATT} \left(C^{src}, \mathbf{s}'_j \right), \\
\mathbf{c}_j &= \left[\mathbf{c}^{mt}_j ; \mathbf{c}^{src}_j \right], \\
\mathbf{s}_j &= \text{GRU}_2 \left(\mathbf{s}'_j, \mathbf{c}_j \right).
\end{aligned}$$

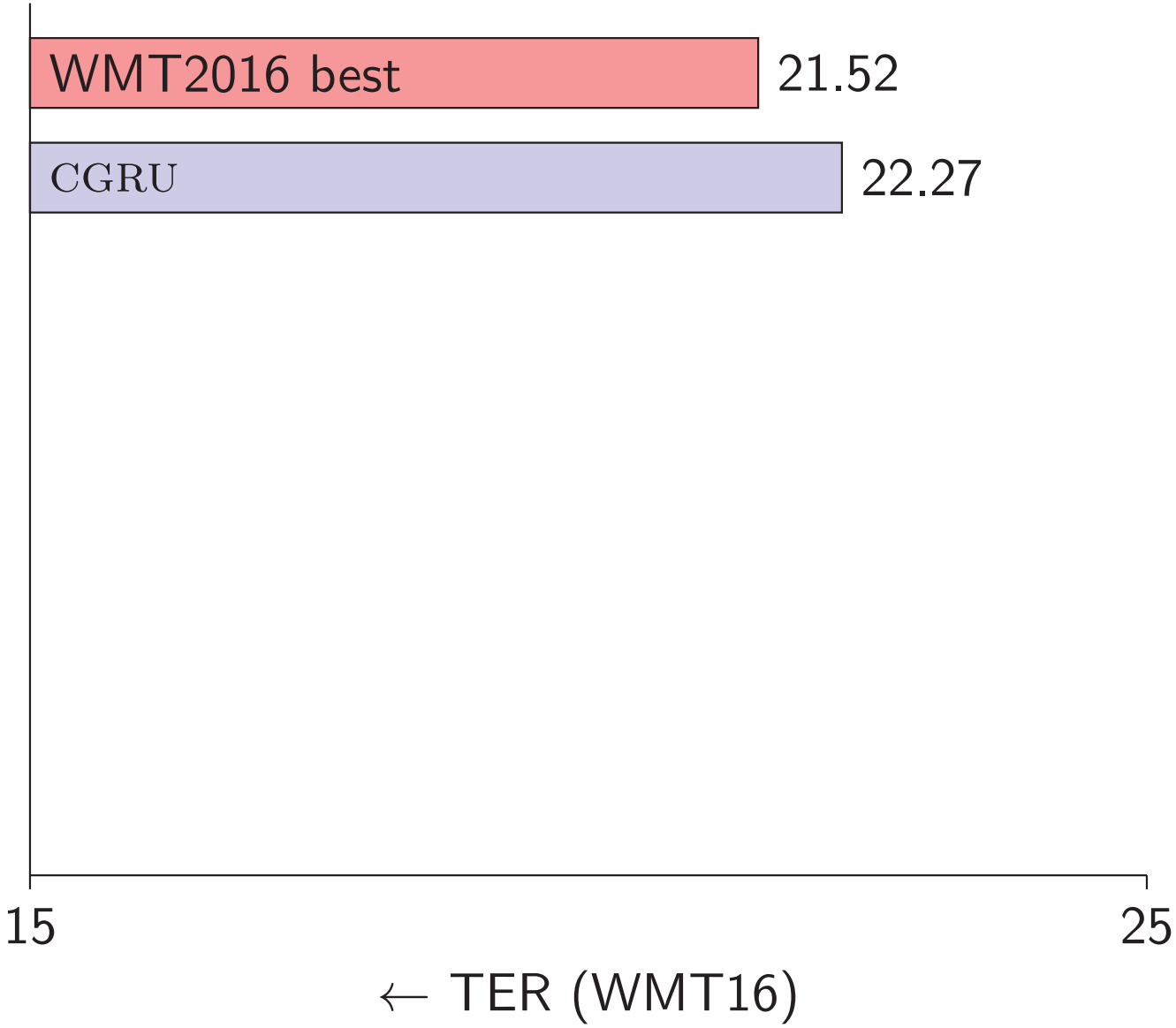
Dual soft attention with hard attention (m-cgru-hard)

$$\mathbf{s}_j = \mathrm{cGRU}_{\text{2-att}}\left(\mathbf{s}_{j-1}, \left[\mathbf{E}[y_{j-1}]; \mathbf{h}^{mt}_{a_j}\right], C^{mt}, C^{src}\right).$$

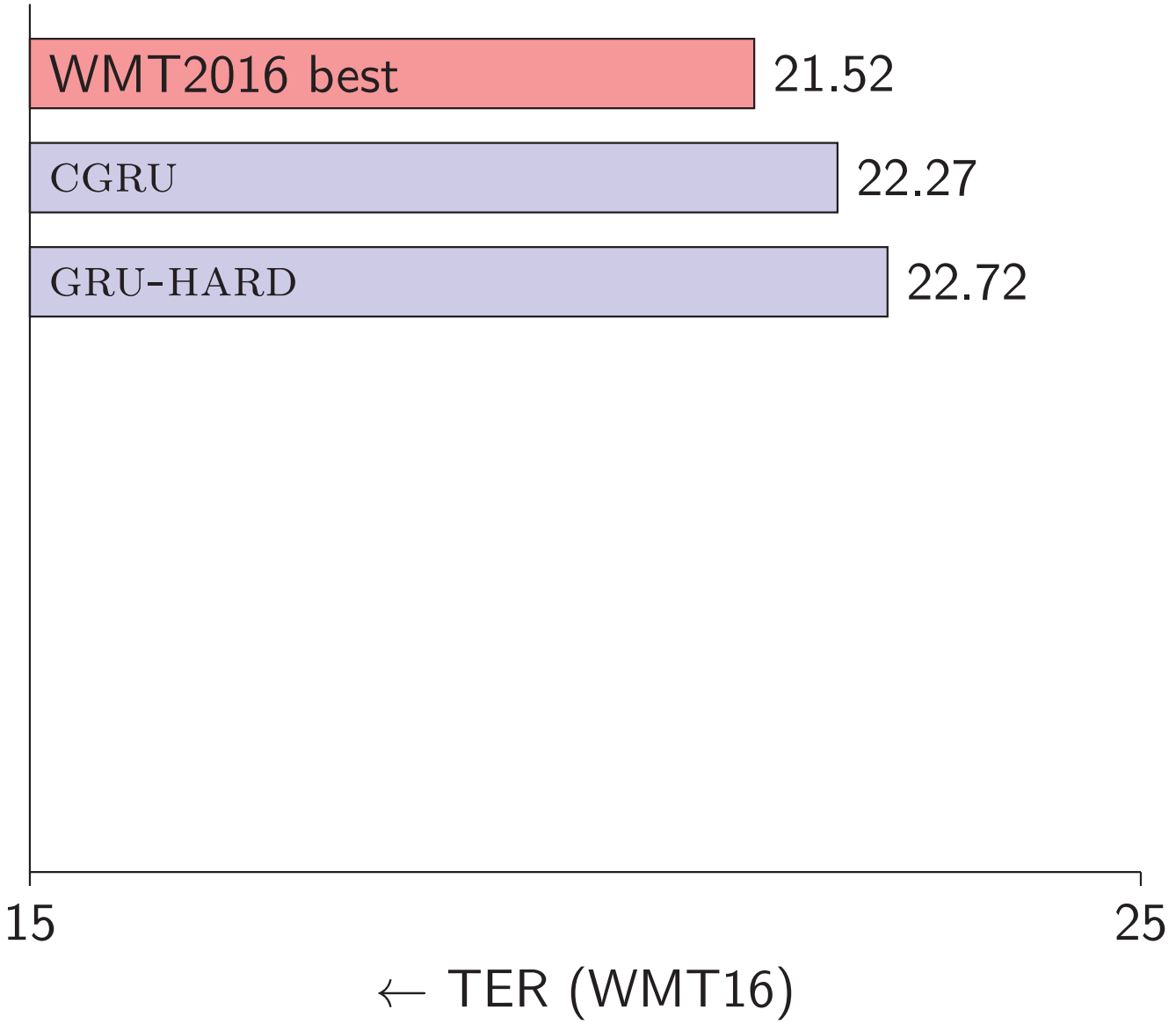
Results



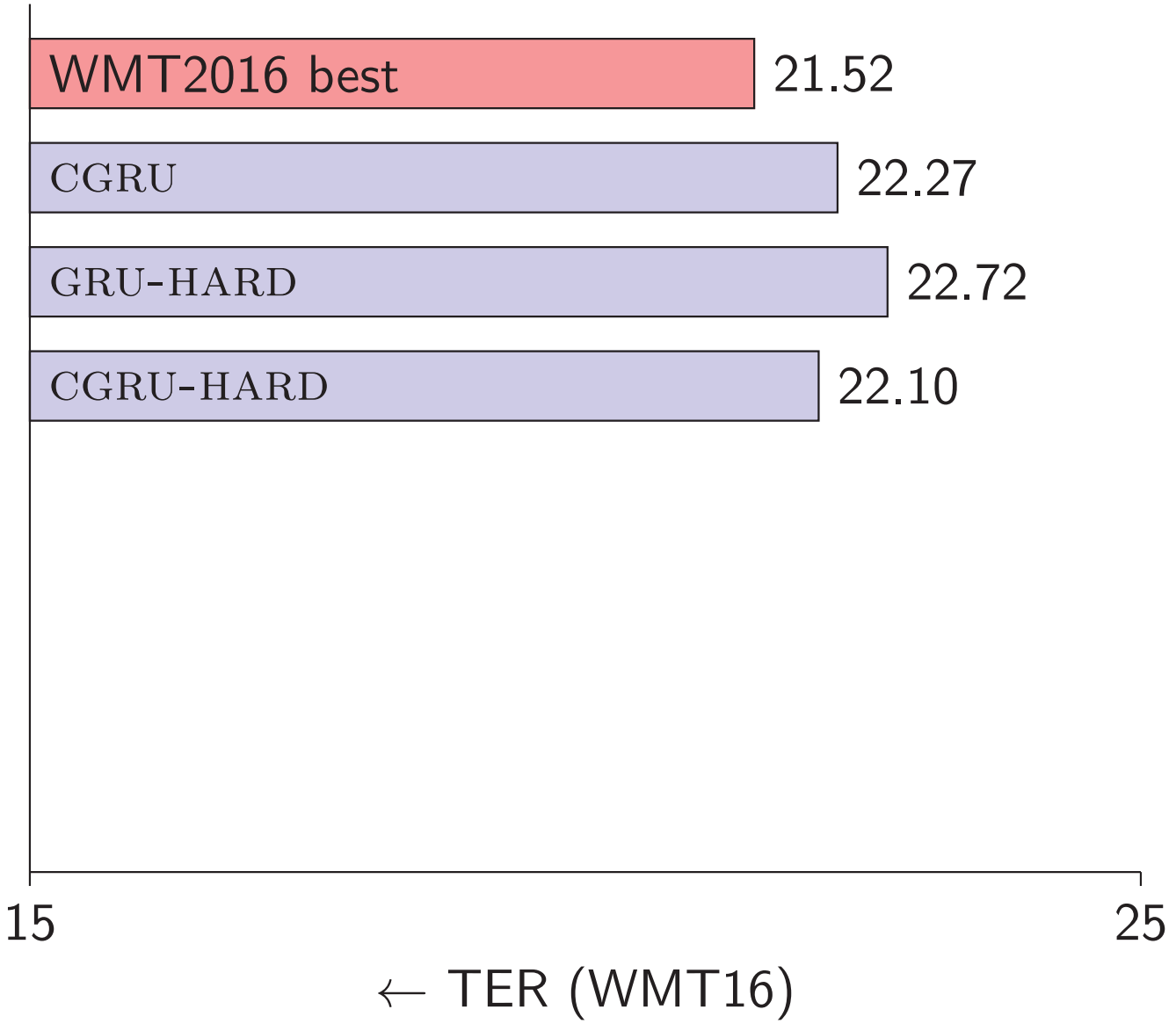
Results



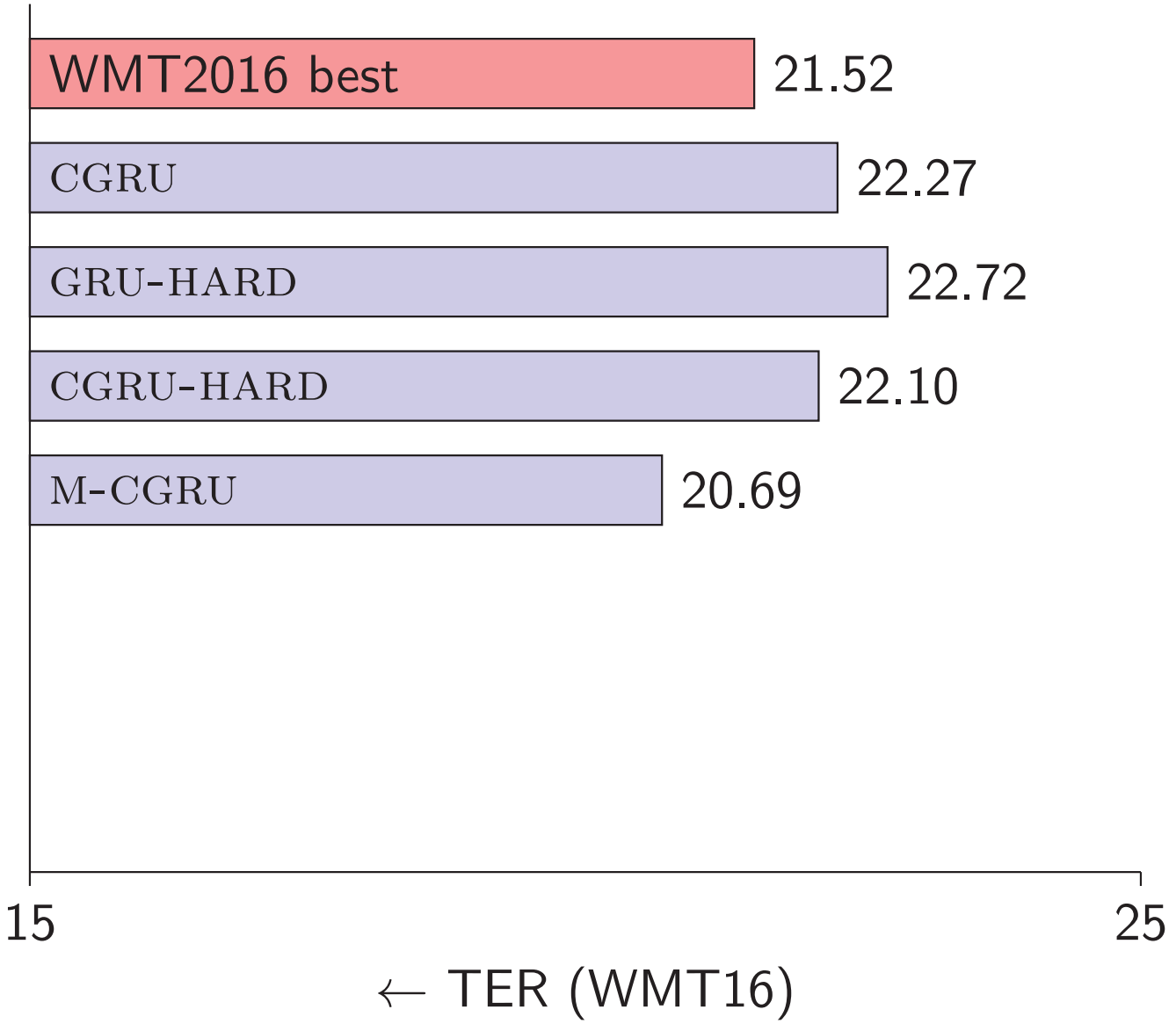
Results



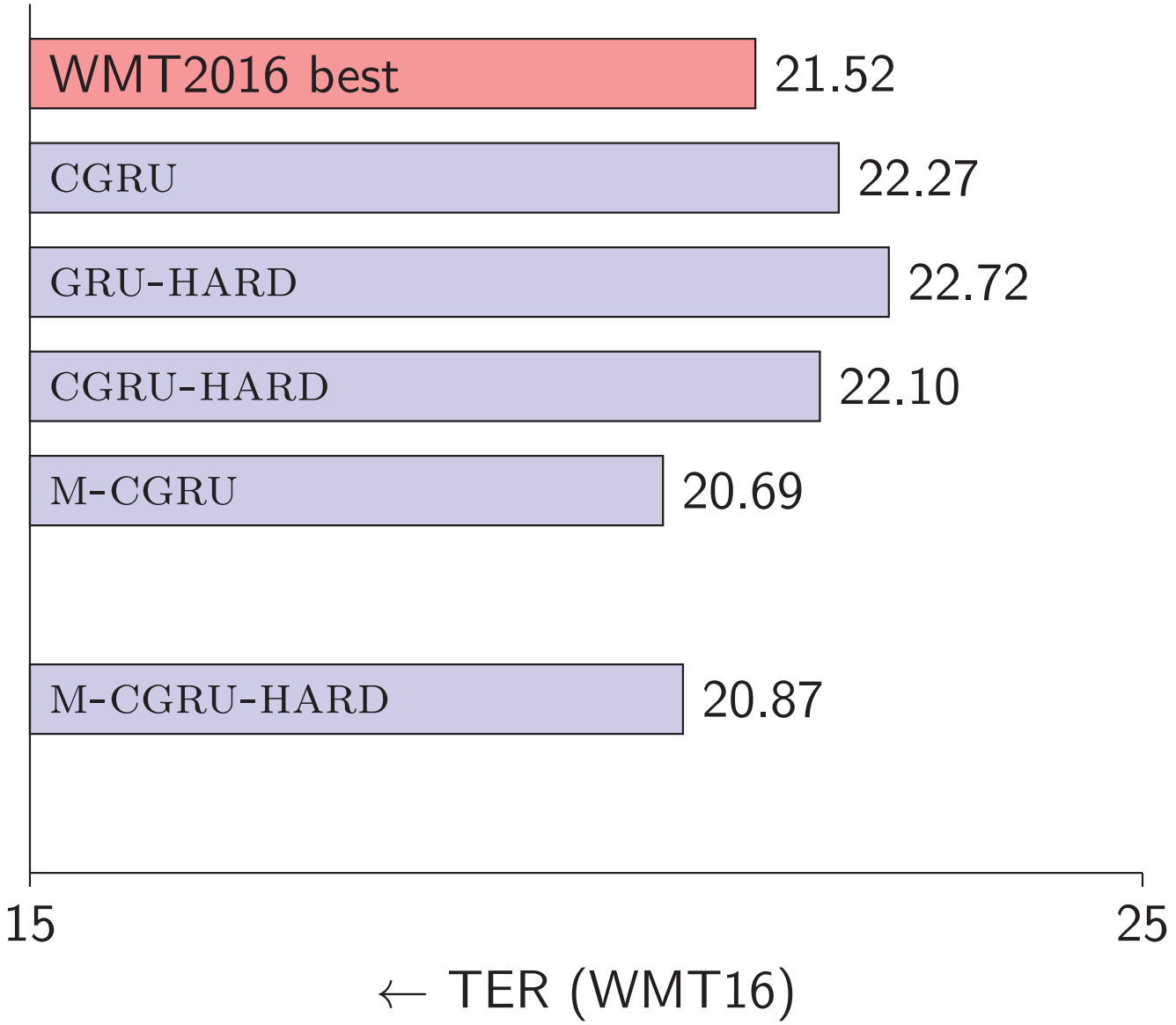
Results


WMT2016 best
21.52
CGRU
22.27
GRU-HARD
22.72
CGRU-HARD
22.10
15
25
← TER (WMT16)

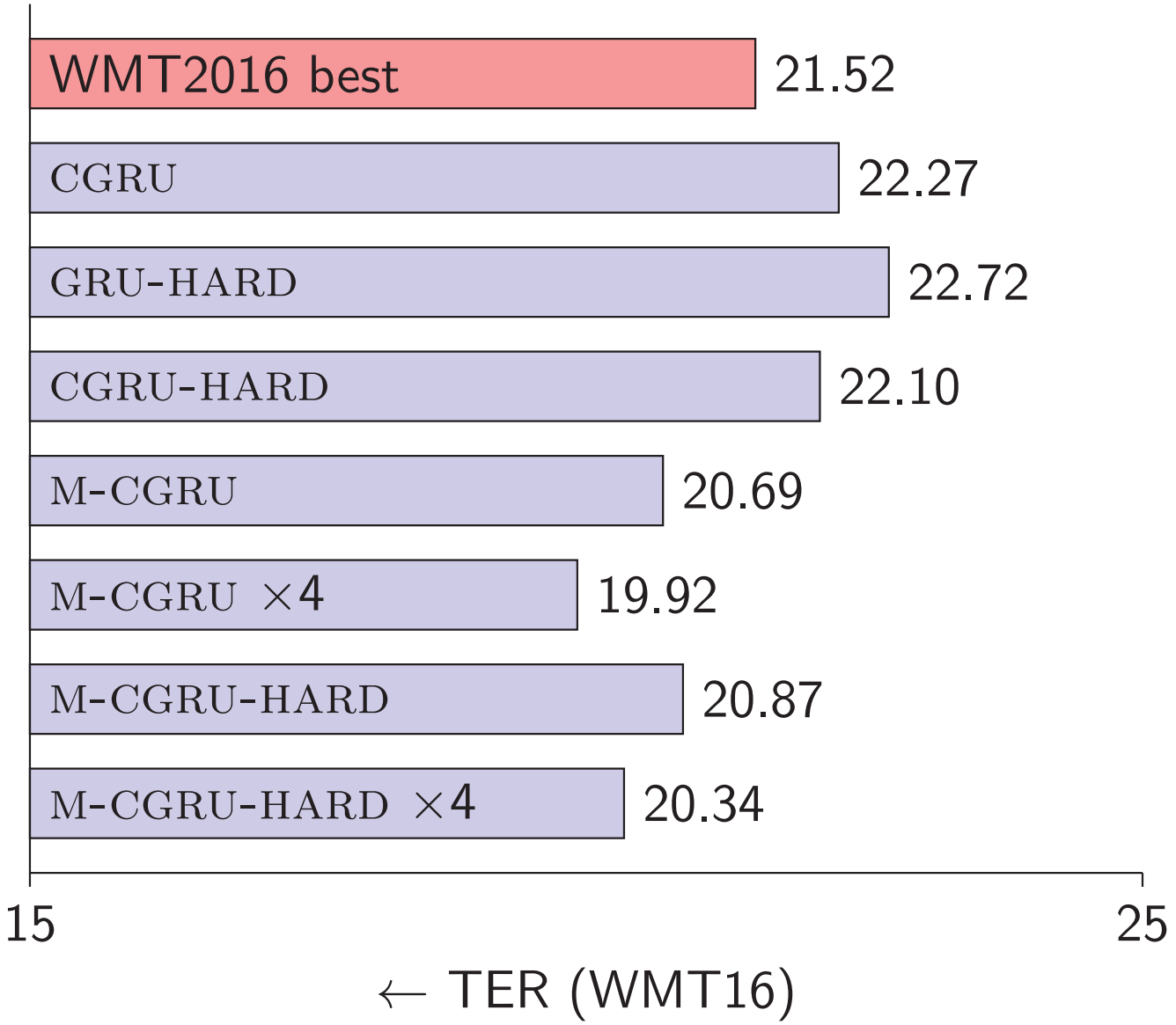
Results


WMT2016 best
21.52
CGRU
22.27
GRU-HARD
22.72
CGRU-HARD
22.10
M-CGRU
20.69
15
25
← TER (WMT16)

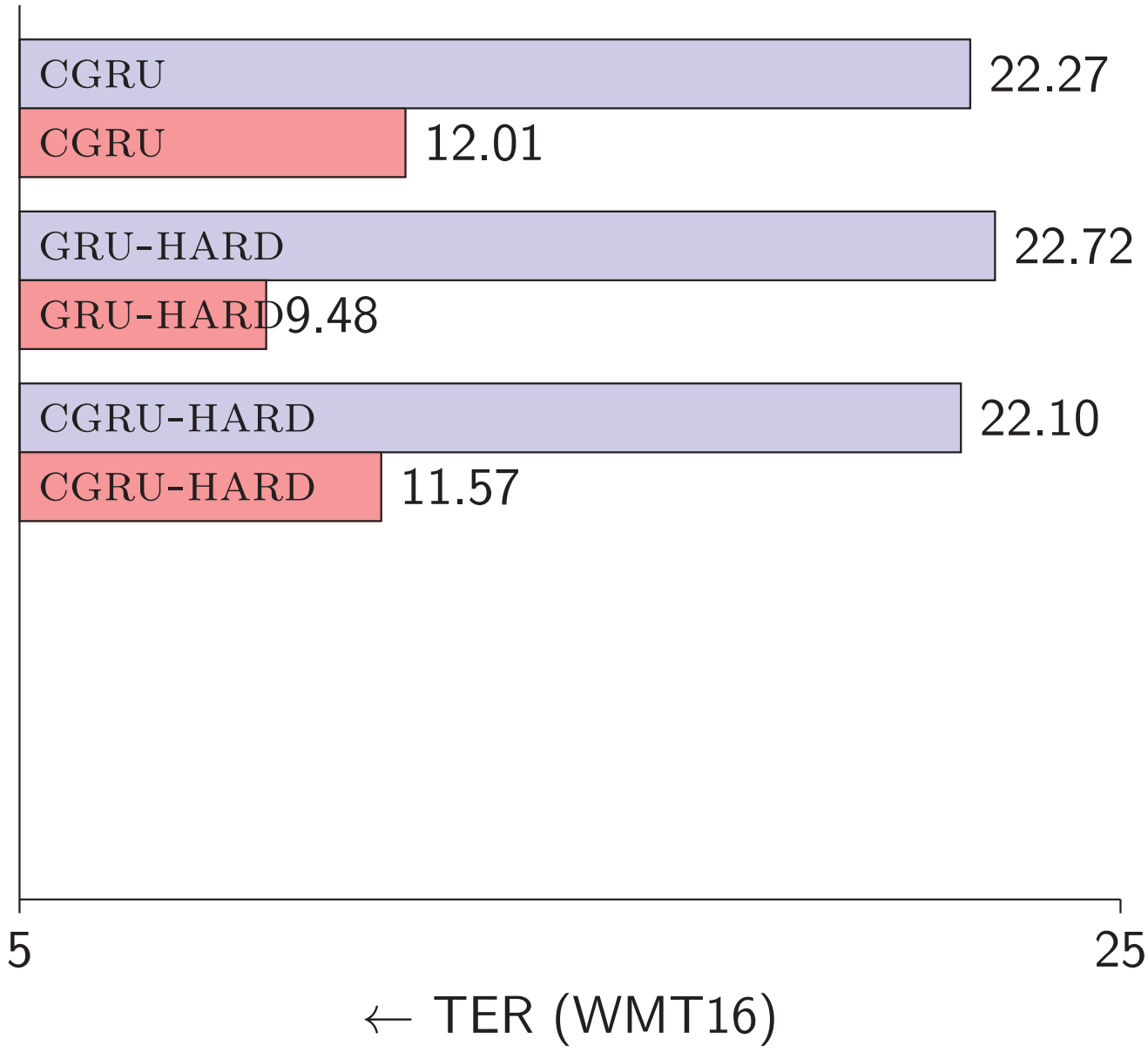
Results


WMT2016 best
21.52
CGRU
22.27
GRU-HARD
22.72
CGRU-HARD
22.10
M-CGRU
20.69
M-CGRU-HARD
20.87
15
25
← TER (WMT16)

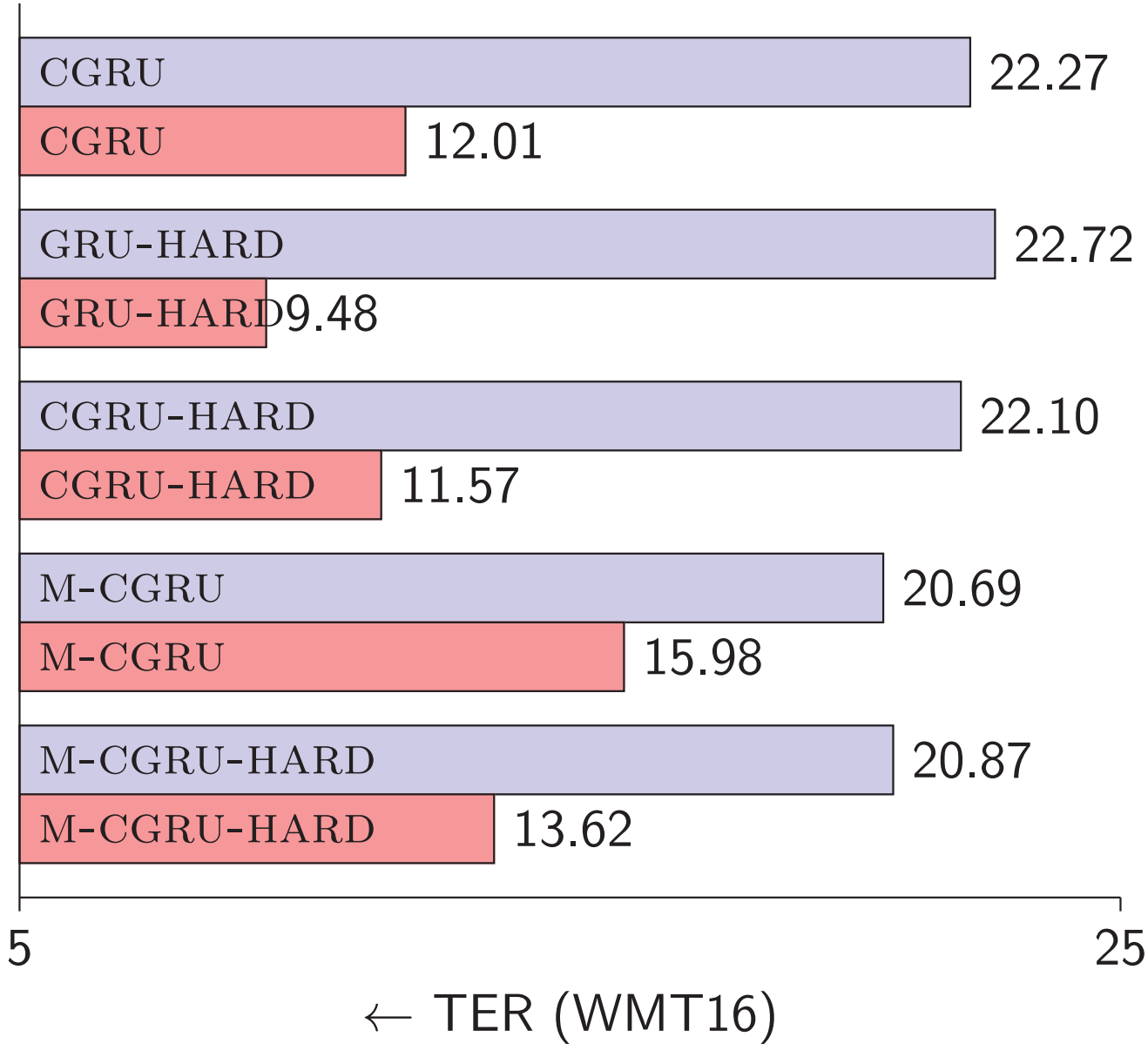
Results


WMT2016 best
21.52
CGRU
22.27
GRU-HARD
22.72
CGRU-HARD
22.10
M-CGRU
20.69
M-CGRU ×4
19.92
M-CGRU-HARD
20.87
M-CGRU-HARD ×4
20.34
15
25
← TER (WMT16)

Faithfulness



Faithfulness



Results for the WMT2017 shared task on APE

Systems	TER	BLEU
FBK EnsembleRerank Primary	19.60	70.07
AMU.multi-transducer-composed PRIMARY	**19.77**	**69.50**
DCU FRANKENAPE-TUNED PRIMARY	20.11	69.19
USAAR NMT-OSM PRIMARY	23.05	65.01
LIG chained syn PRIMARY	23.22	65.12
JXNU JXNU EDITFreq PRIMARY	23.31	65.66
CUNI char conv rnn beam PRIMARY	24.03	64.28
Official Baseline (MT)	24.48	62.49
Baseline 2 (Statistical phrase-based APE)	24.69	62.97

Results for the WMT2017 shared task on APE

#	Ave %	Ave z	System
–	84.8	0.520	Human post edit
1	**78.2**	**0.261**	**AMU**
	77.9	0.261	FBK
	76.8	0.221	DCU
4	73.8	0.115	JXNU
5	71.9	0.038	USAAR
	71.1	0.014	CUNI
	70.2	-0.020	LIG
–	68.6	-0.083	No post edit

Source: WMT 2017 overview paper

Results for the WMT2017 shared task on APE

Systems	Modified	Improved	Deteriorated
FBK Primary	1,607	1,035	334
AMU Primary	**1,583**	**1,040**	**322**
DCU Primary	1,592	1,014	361
...			

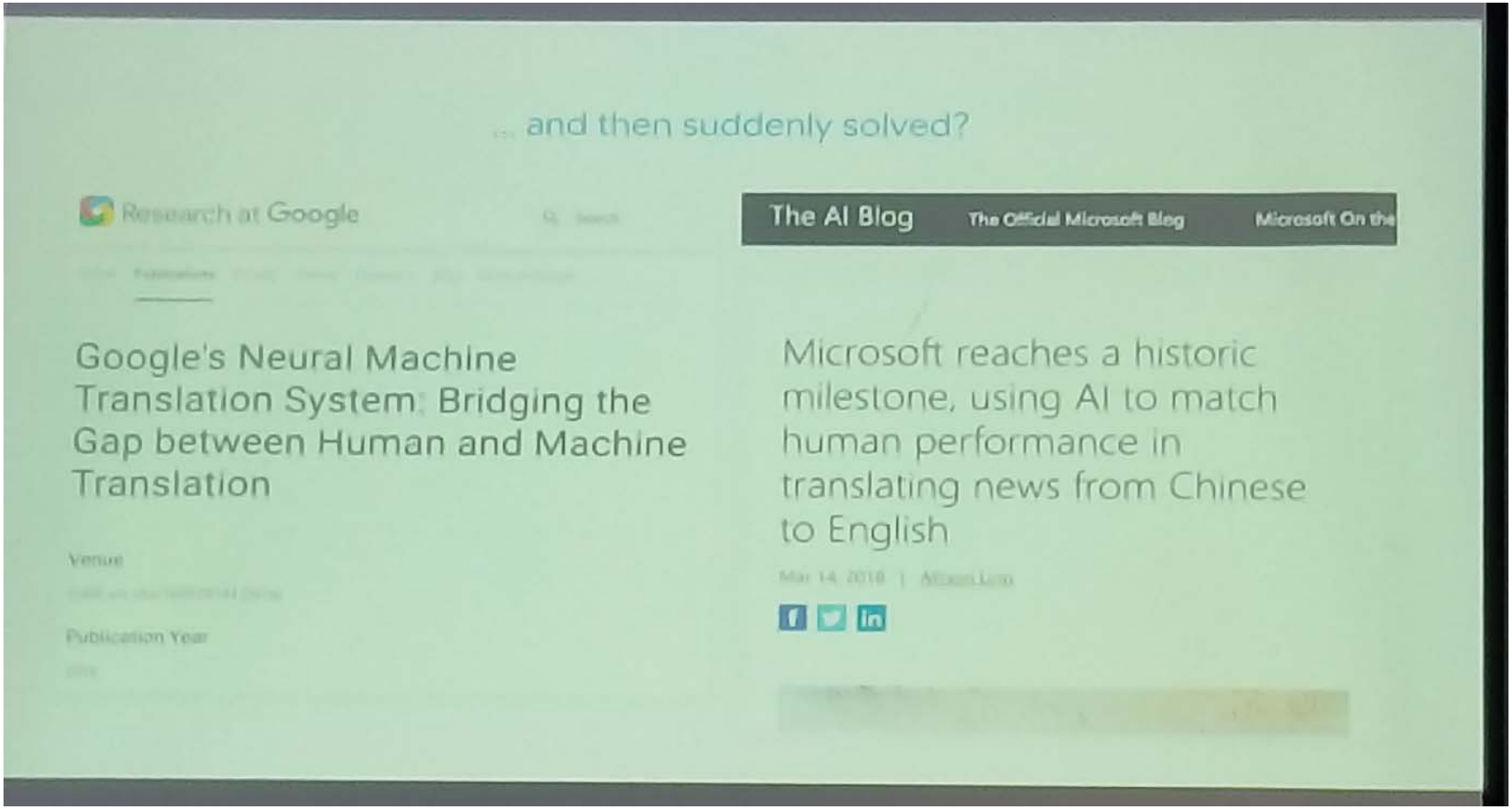
Source: WMT 2017 overview paper

WMT 2018 Shared Task on Automatic post-editing

- First shared tasks to post-edit NMT output (exciting!)
- Still en-de and IT (not ideal!)
- Domain mis-match between artifical data and NMT (bad!)
- More artifical data (good!)

More guesses

General MT is eating your lunch!



Justification

- General QE and APE will be gone before translators even need start worrying;
- QE and APE are bug-fixes that operate within very narrow error margin (too bad to exploit full error margin);
- This error margin might already be gone in many real-word applications.

But ... but... it works, right?

Maybe, maybe not. I think we are mostly seeing:

- Favorably chosen test sets, domains and language pairs;
- Synergy effects (different approaches): SMT+NMT
- System combination effects (similar approches);
- Two-pass decoding effects (see MS results);
- Domain-adaptation or style-transfer effects (**the last hope!**)