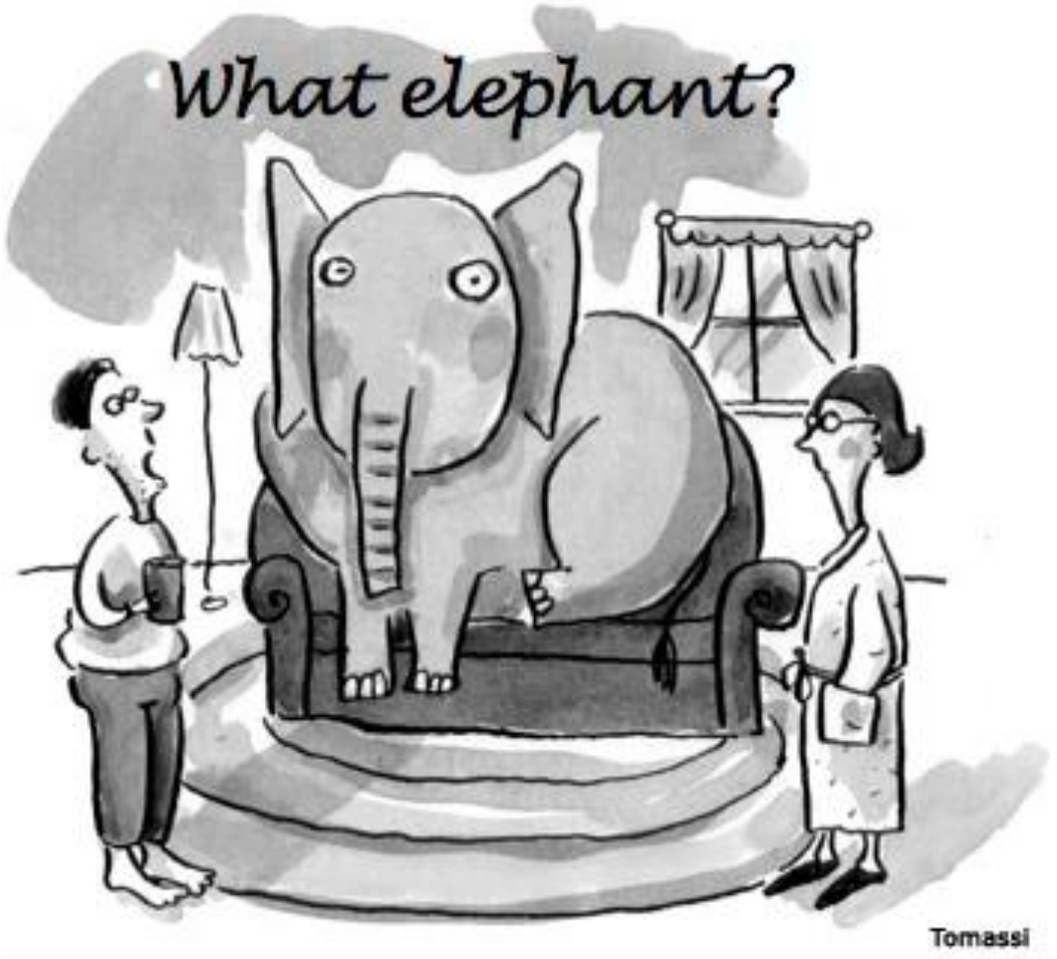
Microsoft

# Practical Neural Machine Translation

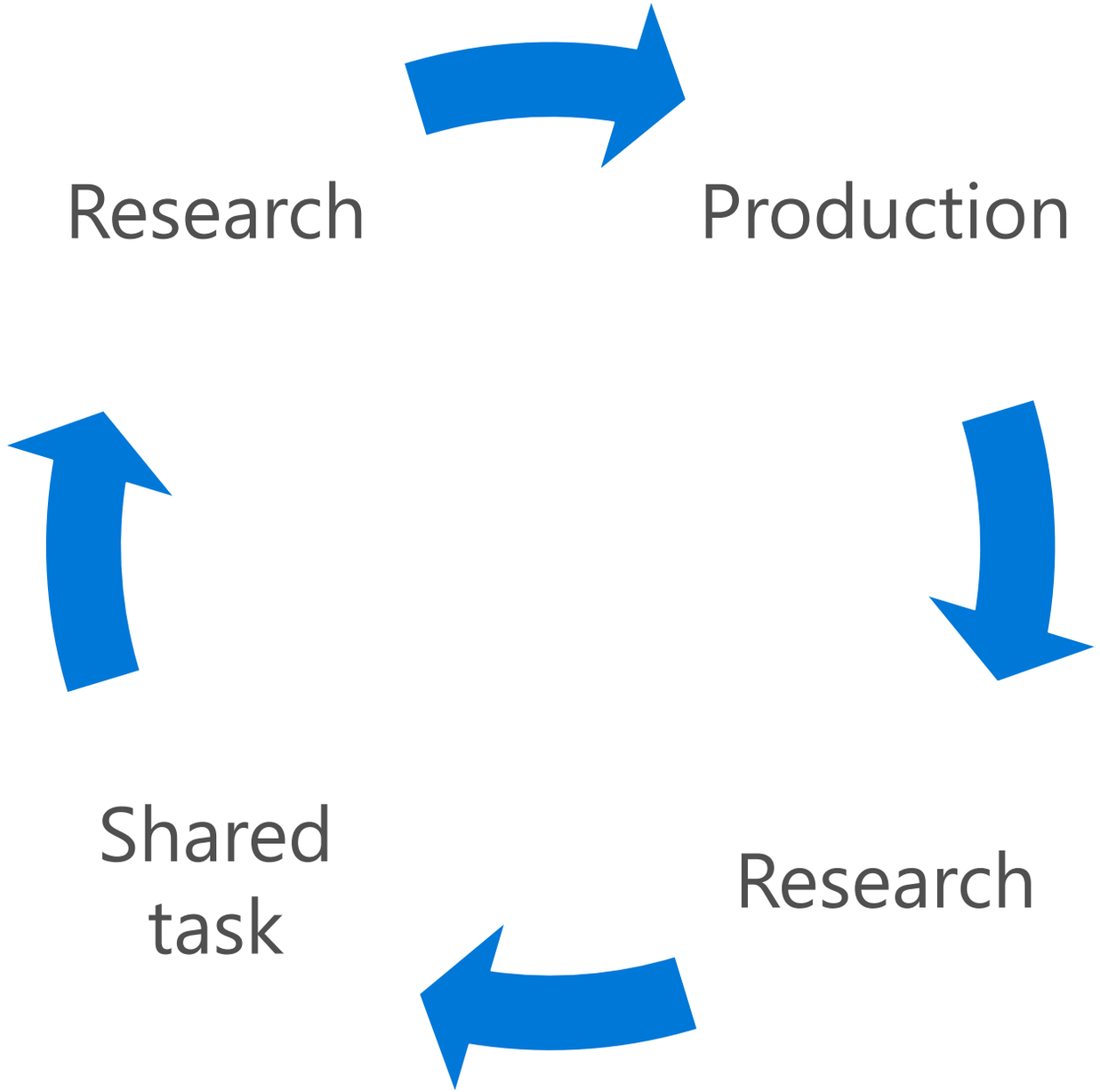
Marcin Junczys-Dowmunt
MT-Class, Pittsburgh

Translator

# About compute

# Shared tasks!


Research
Production
Research
Shared task

# Know your data!
(and fix it)

# Intelligent Selection of Language Model Training Data

**Robert C. Moore** **William Lewis**
Microsoft Research
Redmond, WA 98052, USA
{bobmoore,wilewis}@microsoft.com

# Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora

**Marcin Junczys-Dowmunt**
Microsoft
1 Microsoft Way
Redmond, WA 98121, USA

Cross-Entropy Difference:

$$H_I(x) - H_N(x)$$

Dual Cross-Entropy Score:

$$|H_A(y|x) - H_B(x|y)| + \frac{1}{2}(H_A(y|x) + H_B(x|y))$$

# WMT19 training data – The Good

0.9630 0.7797

Photos: Juanmi Alemany, Henri van der Stelt, Carlos Bernabeu, Jose Augusto Silva y David Acuña.

Fotos: Juanmi Alemany, Henri van der Stelt, Carlos Bernabeu, Jose Augusto Silva y David Acuña.

0.9625 0.2594

Any EU citizen, resident, or an enterprise or association in a Member State, can lodge a complaint with the Ombudsman.

Alle Bürger, Einwohner, Unternehmen oder Verbände in einem Mitgliedstaat können sich beim Bürgerbeauftragten beschweren.

0.9620 0.3164

Other cities are Huambo, Lobito, Lucapa, Benguela, Lubango, Malanje and Cabinda.

Andere Städte sind Huambo, Lobito, Lucapa, Benguela, Lubango, Malanje und Cabinda.

0.9551 1.0000

Turkey is a bridge between Europe and Asia.

Die Türkei ist eine Brücke zwischen Europa und Asien.

# WMT19 training data – The Bad

0.0000 0.0005

Wir hatten 4 Nächte gebucht, sind aber nach 1 Nacht wieder abgereist.

I thoroughly enjoyed my stay there.

0.0000 0.0005

Staff were pleasant and helpful. Room was clean and tidy.

Staff were incredibly friendly and helpful.

0.0000 0.0004

- establish the relative proportions of the components present (mass balance), and - permit the soil residue of concern and to which non-target species are or may be exposed, to

- extractable substances not identified, and - non-extractable residues in soil.

0.0000 0.0004

In the simplest case, a secretive st remy en provence rental is overstepped to an historically isp projecting one of the cards translated above, and the isp uses this utilizatio

Therefore, ip cannot collect thought insignificant totally.

# WMT19 training data – The Ugly

0.0000 1.0000
02/02/2010 Mutually beneficial solution to the Cyprus problem "within reach, ...
07/12/2009 UNO-Generalsekretär Ban Ki-moon: Erklärung zum Internationalen Ta ...

0.0000 1.0000
100 ambitious Lupo drivers give Lupo a warm welcome.
Gemeinsame Weiterfahrt bis nach Wolfsburg, angeführt von einem "Polizei-Lupo".

0.0000 1.0000
100 modern guestrooms with the 4*sup comfort.
Seit 50 Jahren in der gleichen Besitzerfamilie.

0.0000 1.0000
15 rooms and 4 suites with whirlpool or steam bath provide truly high-quality living comfort .
Originalstücke schönster Tiroler Bauernstuben und heimische Handwerkskunst prägen das Interieur.

# WMT18 - Data-Filtering Shared Task

| Dataset | newstest2016 | newstest2017 |
|---|---|---|
| WMT18 | 33.9 | 29.0 |
| Random | 16.2 | 14.1 |
| LangID+Random | 26.6 | 23.3 |
| LangID+Adeq | 35.1 | 30.2 |
| Ablation: no LangID | 15.4 | 12.7 |
| Ablation: no AbsDiff | 33.8 | 29.3 |
| Ablation: no CE-Weight | 31.7 | 27.4 |
| LangID+Adeq+Dom | 36.0 | 31.0 |

# Improving Neural Machine Translation Models with Monolingual Data

**Rico Sennrich** and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk, bhaddow@inf.ed.ac.uk

# Understanding Back-Translation at Scale

**Sergey Edunov**[△] **Myle Ott**[△] **Michael Auli**[△] **David Grangier**[▽*]
[△]Facebook AI Research, Menlo Park, CA & New York, NY.
[▽]Google Brain, Mountain View, CA.

# Effects of noised back-translation

| System | 2016 | 2017 | 2018 |
|---|---|---|---|
| WMT18-Microsoft | 38.6 | 31.3 | 46.5 |
| WMT18-FAIR | - | 32.7 | 44.9 |
| | | | |
| WMT19-baseline | 37.7 | 30.3 | 46.5 |
| + data-filtering | 38.3 | 31.1 | 46.6 |
| + noisy back-translation | 38.9 | 32.8 | 46.3 |
| + fine-tuning | 40.6 | 33.6 | 48.9 |

# Effects of noised back-translation

| System | en | de | both |
|---|---|---|---|
| WMT18-Microsoft | 41.1 | 35.5 | 39.1 |
| WMT18-FAIR | - | - | - |
| | | | |
| WMT19-baseline | 41.8 | 32.5 | 38.2 |
| + data-filtering | 41.7 | 34.0 | 39.0 |
| + noisy back-translation | 38.9 | 40.4 | 39.7 |
| + fine-tuning | 42.2 | 39.2 | 41.2 |

# Neural Machine Translation of Rare Words with Subword Units

**Rico Sennrich** and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk, bhaddow@inf.ed.ac.uk

# SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing

**Taku Kudo** **John Richardson**
Google, Inc.
{taku,johnri}@google.com

# Applying SentencePiece

Obama receives Netanyahu

The relationship between Obama and Netanyahu is not exactly friendly.

The two wanted to talk about the implementation of the international agreement and about Teheran's destabilising activities in the Middle East.

The meeting was also planned to cover the conflict with the Palestinians and the disputed two state solution.

Relations between Obama and Netanyahu have been strained for years.

▁Obama ▁receives ▁Netanyahu

▁The ▁relationship ▁between ▁Obama ▁and ▁Netanyahu ▁is ▁not ▁exactly ▁friendly .

▁The ▁two ▁wanted ▁to ▁talk ▁about ▁the ▁implementation ▁of ▁the ▁international ▁agreement ▁and ▁about ▁Teheran ' s ▁de stabil ising ▁activities ▁in ▁the ▁Middle ▁East .

▁The ▁meeting ▁was ▁also ▁planned ▁to ▁cover ▁the ▁conflict ▁with ▁the ▁Palestinians ▁and ▁the ▁dispute d ▁two ▁state ▁solution .

▁Relations ▁between ▁Obama ▁and ▁Netanyahu ▁have ▁been ▁strain ed ▁for ▁years .

# BPE versus SentencePiece

```
cat newstest2016.en-de.en | \
./tokenizer.perl -l en -a | \
./apply_bpe.py -c bpe.ende | \
./marian-decoder -c config.yml | \
sed 's/\@\@ //g' | \
./detokenizer.perl -l de > newstest2016.en-de.out
```

# BPE versus SentencePiece

```
cat newstest2016.en-de.en | \
./marian-decoder -c config.yml > newstest2016.en-de.out
```

# BPE versus SentencePiece

| BPE | SentencePiece |
|---|---|
| Poli@@ zei\|ch@@ ef | _Polizei chef |
| ver@@ hän@@ g\|nis@@ vollen | _ver h äng nis vollen |
| Universit@@ ä@@ t\|s @-@ Mitarbeiter | _Universität s - Mitarbeiter |
| Schie@@ ß\|en | _Schieß en |
| be\|su@@ cht\|en | _besucht en |
| auf@@ gere@@ g\|t | _a uf\|gereg\|t |
| Be@@ urlau@@ b\|ung | _Be urlaub ung |

# Build strong baselines!

(don’t lie to yourself)

# Six Challenges for Neural Machine Translation

**Philipp Koehn**
Computer Science Department
Johns Hopkins University
phi@jhu.edu

**Rebecca Knowles**
Computer Science Department
Johns Hopkins University
rknowles@jhu.edu

# Six challenges for Neural Machine Translation

1. Low quality for out of domain-input
2. Lower quality than SMT for low-resource settings
3. Worse for rare words in inflectional languages
4. Lower quality for sentences above 60 tokens
5. Attention is not word alignment
6. Beam-search deteriorates with larger beams

# Stronger Baselines for Trustable Results in Neural Machine Translation

**Michael Denkowski**
Amazon.com, Inc.
mdenkows@amazon.com

**Graham Neubig**
Carnegie Mellon University
gneubig@cs.cmu.edu

# Stronger baselines

"New research regularly introduces architectural and algorithmic improvements that lead to significant gains over 'vanilla' NMT implementations. However, these new techniques are rarely evaluated in the context of previously published techniques, specifically those that are widely used in state-of-the art production and shared-task systems. As a result, it is often difficult to determine whether improvements from research will carry over to systems deployed for real-world use."

# Stronger baselines

- Training using Adam with multiple restarts and learning rate annealing
- Sub-word translation via byte pair encoding
- Decoding with ensembles of independently trained models

# Revisiting Low-Resource Neural Machine Translation: A Case Study

**Rico Sennrich**[1,2] **Biao Zhang**[1]

[1]School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk, B.Zhang@ed.ac.uk

[2]Institute of Computational Linguistics, University of Zurich

# Choose your model!
(and your toolkit)

# Attention Is All You Need

**Ashish Vaswani**[*]
Google Brain
avaswani@google.com

**Noam Shazeer**[*]
Google Brain
noam@google.com

**Niki Parmar**[*]
Google Research
nikip@google.com

**Jakob Uszkoreit**[*]
Google Research
usz@google.com

**Llion Jones**[*]
Google Research
llion@google.com

**Aidan N. Gomez**[*] [†]
University of Toronto
aidan@cs.toronto.edu

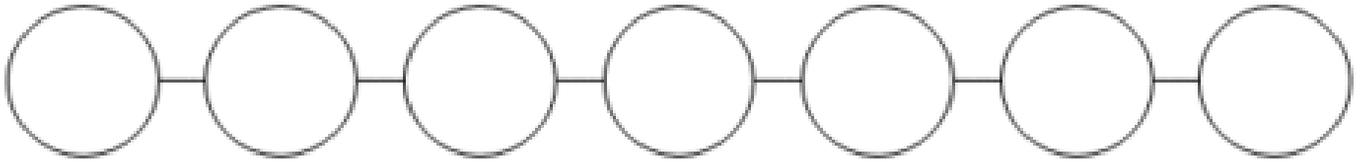
**Łukasz Kaiser**[*]
Google Brain
lukaszkaiser@google.com

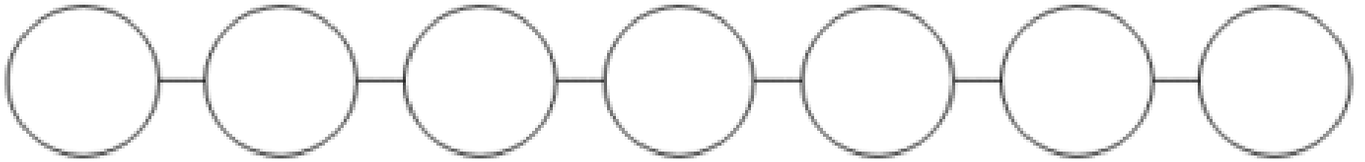
**Illia Polosukhin**[*] [‡]
illia.polosukhin@gmail.com

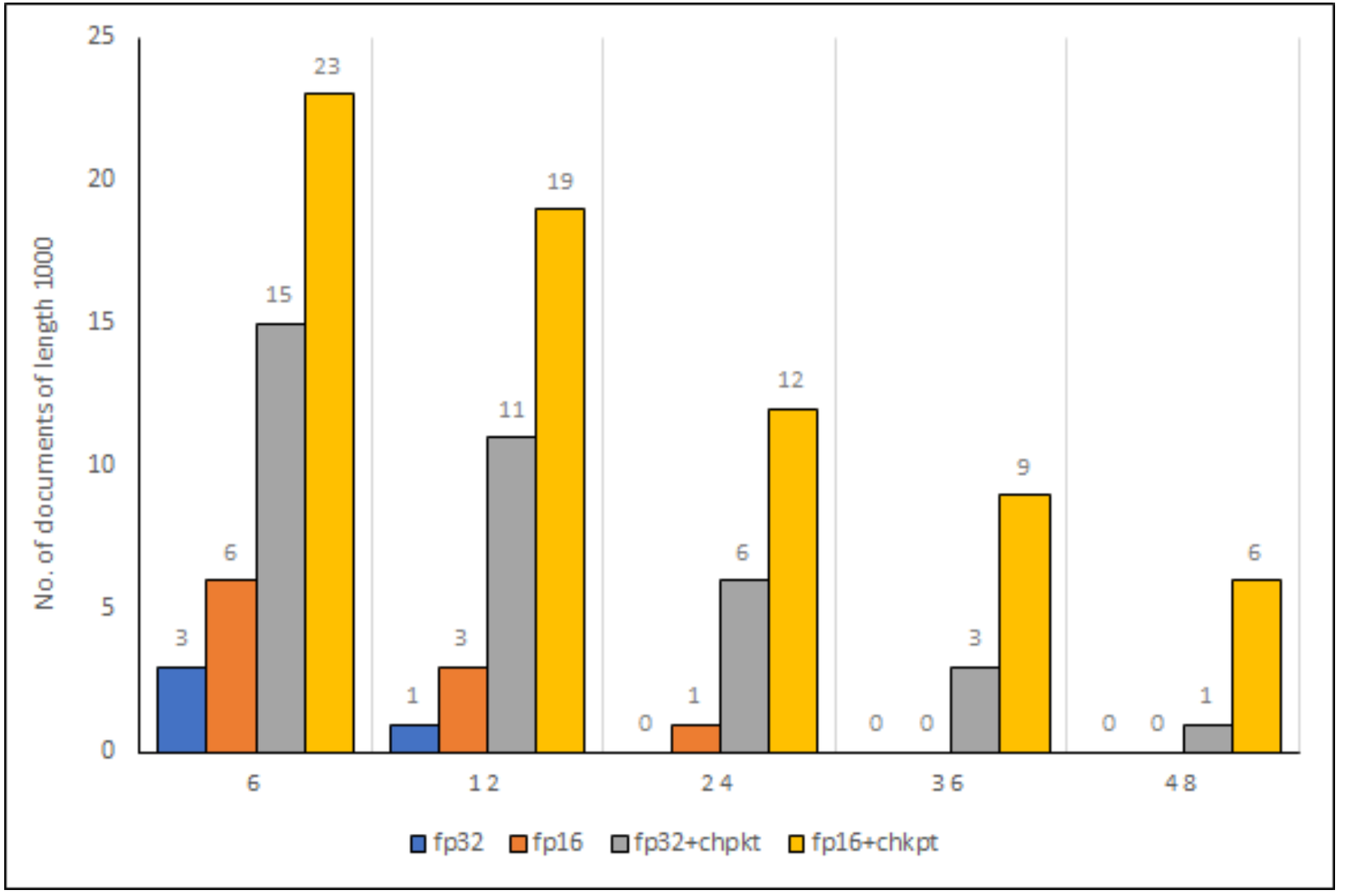
# Marian: Fast Neural Machine Translation in C++

**Marcin Junczys-Dowmunt**[†] **Roman Grundkiewicz**[*‡] **Tomasz Dwojak**[*]
**Hieu Hoang** **Kenneth Heafield**[‡] **Tom Neckermann**[‡]
**Frank Seide**[†] **Ulrich Germann**[‡] **Alham Fikri Aji**[‡]
**Nikolay Bogoychev**[‡] **André F. T. Martins**[¶] **Alexandra Birch**[‡]

[†]Microsoft Translator [*]Adam Mickiewicz University in Poznań
[‡]University of Edinburgh [¶]Unbabel






No. of documents of length 1000
25
20
15
10
5
0
23
19
15
12
11
9
6
6
6
3
3
3
1
1
0
1
0
0
0
0
1
6
6
12
24
36
48
fp32
fp16
fp32+chpkt
fp16+chkpt

# New slide on initialization of deep transformer via depth-scaling. Discovered independently in:

- Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation
  Marcin Junczys-Dowmunt
  https://arxiv.org/abs/1907.06170

- Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention
  Biao Zhang, Ivan Titov, Rico Sennrich
  https://arxiv.org/abs/1908.11365

# Automatic evaluation

(it works just barely)

# A Call for Clarity in Reporting BLEU Scores

**Matt Post**
Amazon Research
Berlin, Germany

```
./sacrebleu.py -t wmt16 -l ende --echo src | \
./marian-decoder -c config.yml | \
./sacrebleu.py -t wmt16 -l ende
```

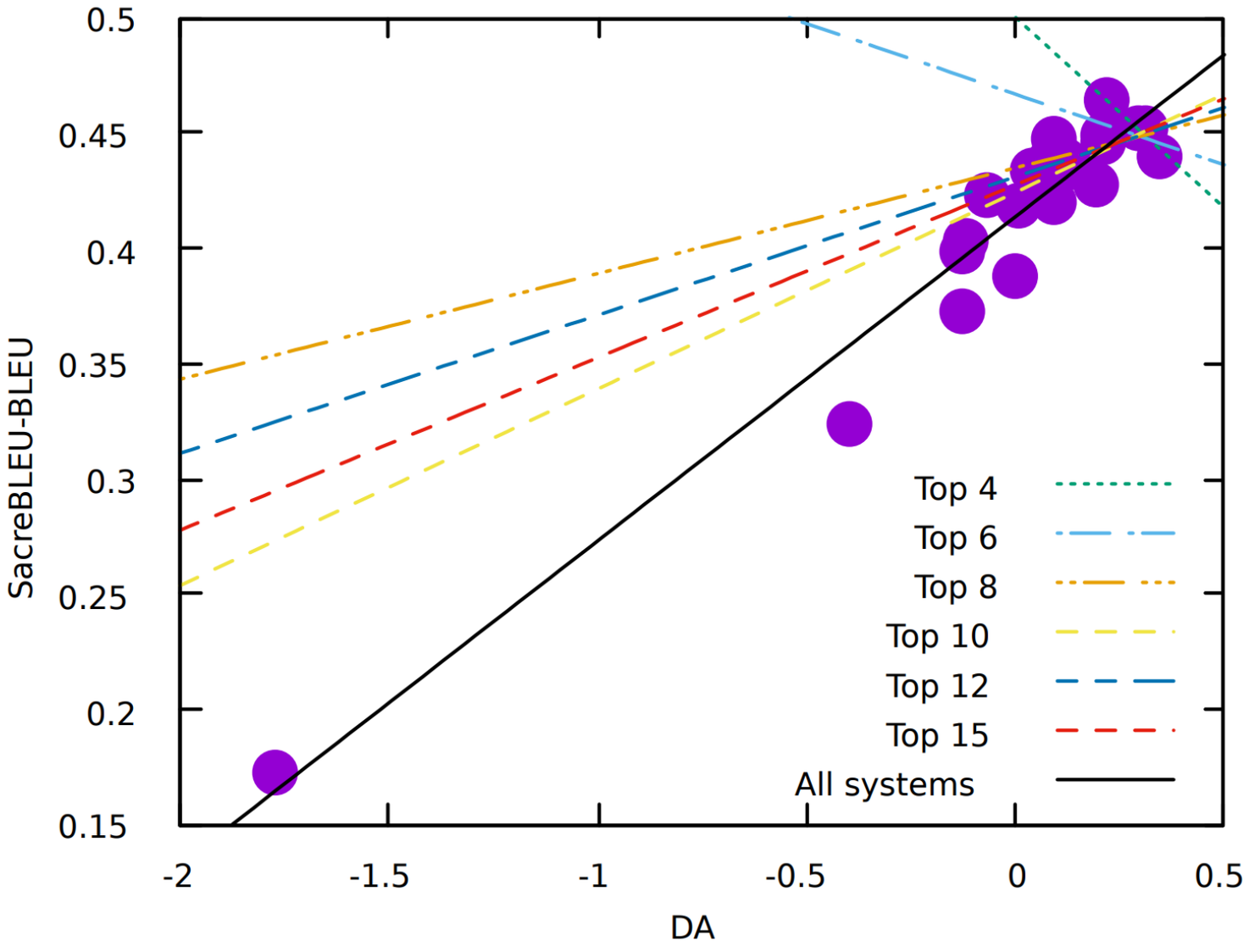
# Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

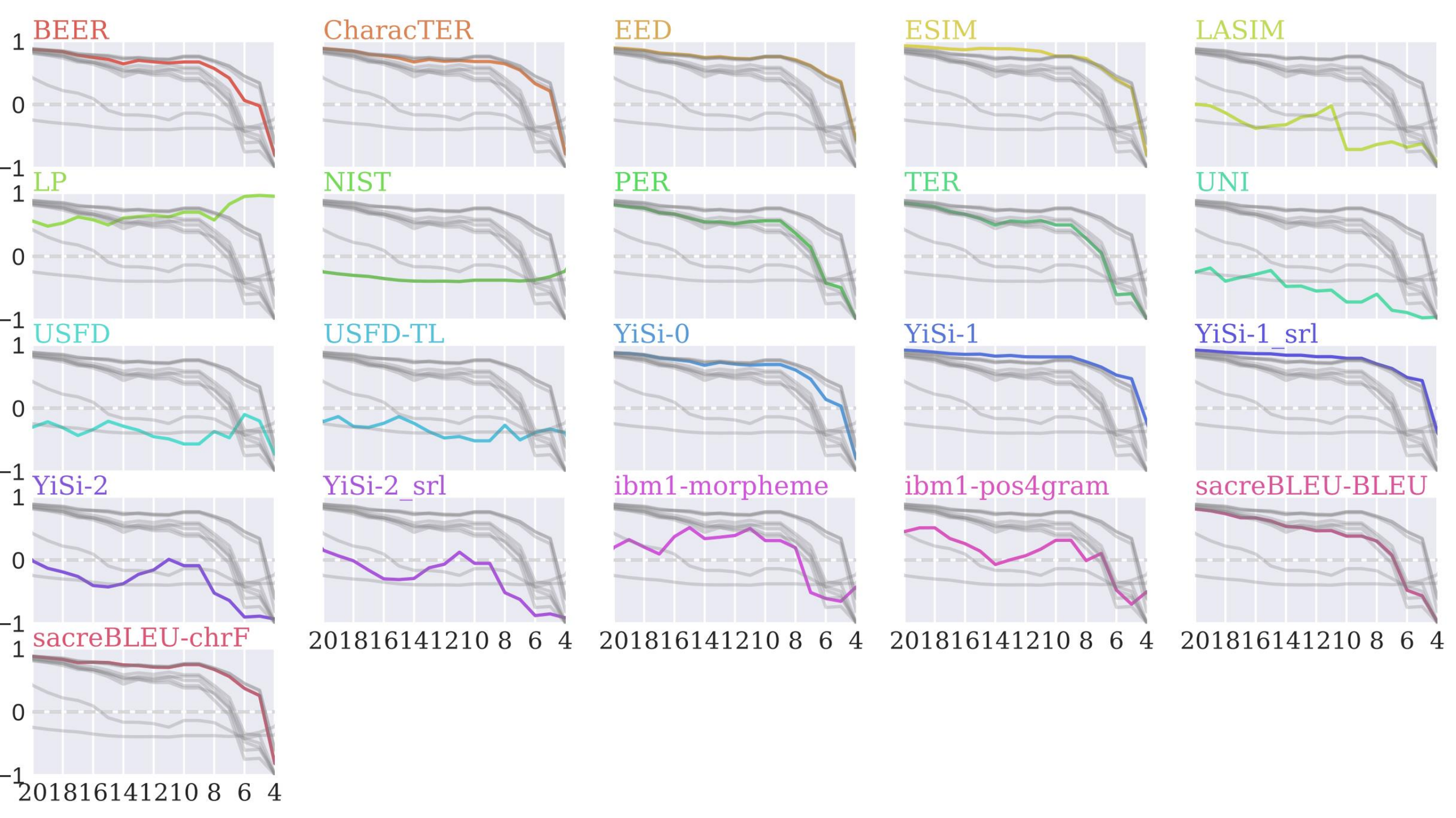
**Qingsong Ma**
Tencent-CSIG, AI Evaluation Lab
qingsong.mqs@gmail.com

**Johnny Tian-Zheng Wei**
UMass Amherst, CICS
jwei@umass.edu

**Ondřej Bojar**
Charles University, MFF ÚFAL
bojar@ufal.mff.cuni.cz

**Yvette Graham**
Dublin City University, ADAPT
graham.yvette@gmail.com


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
SacreBLEU-BLEU
-2
-1.5
-1
-0.5
0
0.5
DA
Top 4
Top 6
Top 8
Top 10
Top 12
Top 15
All systems


BEER
CharacTER
EED
ESIM
LASIM
LP
NIST
PER
TER
UNI
USFD
USFD-TL
YiSi-0
YiSi-1
YiSi-1_srl
YiSi-2
YiSi-2_srl
ibm1-morpheme
ibm1-pos4gram
sacreBLEU-BLEU
sacreBLEU-chrF
1
0
−1
20 18 16 14 12 10 8 6 4

# New slide: also worth reading

On The Evaluation of Machine Translation Systems Trained With Back-Translation

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, Michael Auli

https://arxiv.org/abs/1908.05204

# Human parity!(?)

# Achieving Human Parity on Automatic Chinese to English News Translation

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, and Ming Zhou

Microsoft AI & Research

# Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

**Samuel Läubli**[1] **Rico Sennrich**[1,2] **Martin Volk**[1]

[1]Institute of Computational Linguistics, University of Zurich
{laeubli,volk}@cl.uzh.ch

[2]School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk

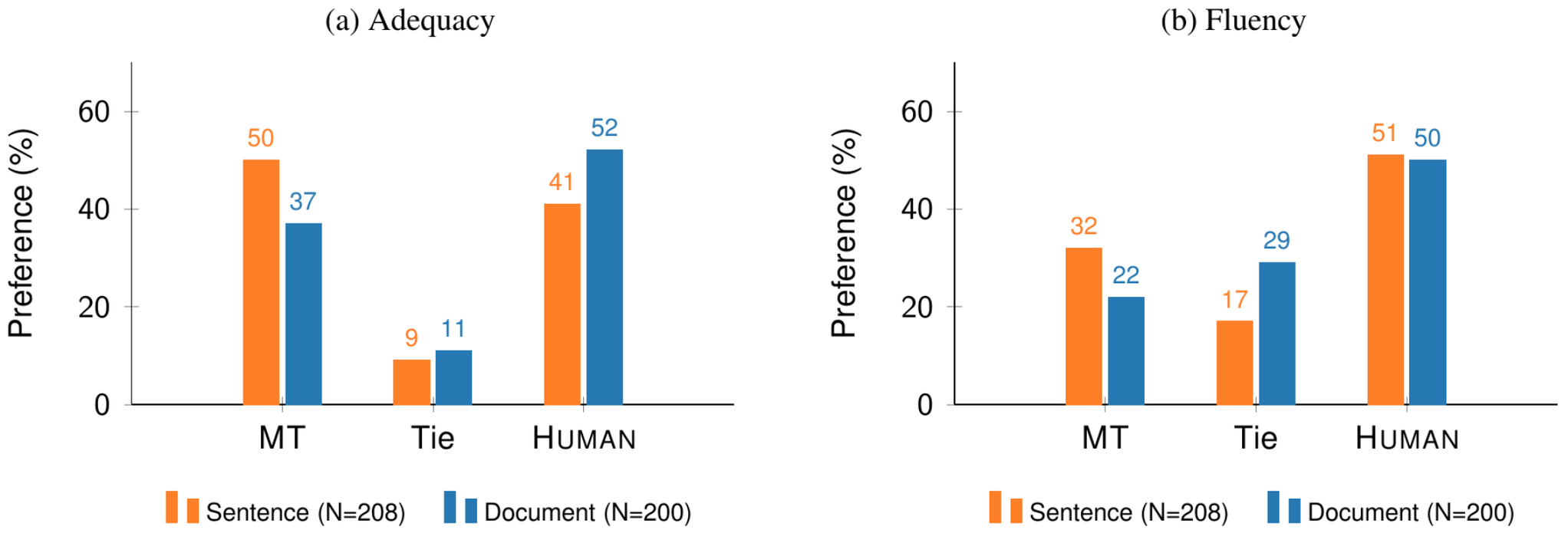



Figure 1: Raters prefer human translation more strongly in entire documents. When evaluating isolated sentences in terms of adequacy, there is no statistically significant difference between HUMAN and MT; in all other settings, raters show a statistically significant preference for HUMAN.

# Findings of the 2019 Conference on Machine Translation (WMT19)

**Loïc Barrault**
Le Mans Université

**Ondřej Bojar**
Charles University

**Marta R. Costa-jussà**
UPC

**Christian Federmann**
Microsoft Cloud + AI

**Mark Fishel**
University of Tartu

**Yvette Graham**
Dublin City University

**Barry Haddow**
University of Edinburgh

**Matthias Huck**
LMU Munich

**Philipp Koehn**
JHU / University of Edinburgh

**Shervin Malmasi**
Harvard Medical School

**Christof Monz**
University of Amsterdam

**Mathias Müller**
University of Zurich

**Santanu Pal**
Saarland University

**Matt Post**
JHU

**Marcos Zampieri**
University of Wolverhampton

**English→German**

| Ave. | Ave. z | System |
|---|---|---|
| 90.3 | 0.347 | Facebook-FAIR |
| 93.0 | 0.311 | Microsoft-WMT19-sent-doc |
| 92.6 | 0.296 | Microsoft-WMT19-doc-level |
| 90.3 | 0.240 | HUMAN |
| 87.6 | 0.214 | MSRA-MADL |
| 88.7 | 0.213 | UCAM |
| 89.6 | 0.208 | NEU |
| 87.5 | 0.189 | MLLP-UPV |
| 87.5 | 0.130 | eTranslation |
| 86.8 | 0.119 | dfki-nmt |
| 84.2 | 0.094 | online-B |
| 86.6 | 0.094 | Microsoft-WMT19-sent-level |
| 87.3 | 0.081 | JHU |
| 84.4 | 0.077 | Helsinki-NLP |
| 84.2 | 0.038 | online-Y |
| 83.7 | 0.010 | lmu-ctx-tf-single |
| 84.1 | 0.001 | PROMT-NMT |
| 82.8 | −0.072 | online-A |
| 82.7 | −0.119 | online-G |
| 80.3 | −0.129 | UdS-DFKI |
| 82.4 | −0.132 | TartuNLP-c |
| 76.3 | −0.400 | online-X |
| 43.3 | −1.769 | en-de-task |

# Scale up!

# BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

**Jacob Devlin** **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

# Language Models are Unsupervised Multitask Learners

**Alec Radford** [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

# English Input – Document-level

<BEG> In 911 Call, Professor Admits to Shooting Girlfriend<SEP> In a 911 call, his voice only slightly shaky, college professor Shannon Lamb told police he had shot his girlfriend and officers needed to get over to their house.<SEP> Lamb made a point to say his "sweet dog" was there alive and probably upset, and said the dead woman's family contacts could be found on her phone.<SEP> Inside the home, officers found Amy Prentiss' body and a hand-written note scribbled on a white legal pad: "I am so very sorry I wish I could take it back I loved Amy and she is the only woman who ever loved me," read the letter authorities say was signed by Lamb.<SEP> There was no indication that Lamb, who was teaching two online classes for Delta State University in Cleveland, Mississippi, had already traveled 300 miles to the school's campus, where police believe he shot and killed a well-liked history professor, Ethan Schmidt, in the doorway to his office.<SEP> Delta State University police chief Lynn Buford said university officials heard about the shooting at 10:18 a.m.<SEP> He said Lamb made the fateful 911 call sometime after that.<SEP> By the end of the day, there would be one more death: Lamb took his own life as police closed in on him.<SEP> A day after the school shooting forced students and faculty to hide behind locked doors, authorities were still trying to piece together what motivated Lamb.<SEP> The details released by investigators at both ends of the state as well as students and staff who knew him helped paint a picture of a talented but possibly troubled teacher.<SEP> Students said they looked forward to his class.<SEP> Police in Gautier, where Prentiss died, said he had no history of violence or criminal record.<SEP> Schmidt himself had included Lamb in a book he wrote where he acknowledged the "wonderful people" he shared his academic life with.<SEP> Both taught in the Division of Social Sciences and History, which lists 17 faculty members, and many students took courses from both.<SEP> At the same time, there were some inclinations of problems.<SEP><BRK>

# German Translation – Document-level

<BEG> In 911-Anruf gibt Professor zu, Freundin erschossen zu haben<SEP> In einem 911-Anruf sei seine Stimme nur leicht wackelig gewesen, sagte College-Professor Shannon Lamb der Polizei, er habe seine Freundin erschossen und Beamte hätten in ihr Haus eindringen müssen.<SEP> Lamb gab zu Protokoll, sein "süßer Hund" sei lebend da gewesen und vermutlich aufgebracht gewesen und sagte, die Familienkontakte der toten Frau seien auf ihrem Handy gefunden worden.<SEP> Im Inneren des Hauses fanden Beamte die Leiche von Amy Prentiss und einen handschriftlichen Zettel, der auf ein weißes Polster gekritzelt war: "Es tut mir so leid, dass ich wünschte, ich könnte sie zurücknehmen, ich liebte Amy und sie ist die einzige Frau, die mich jemals geliebt hat", hieß es in dem Brief, den die Behörden unterschrieben hatten.<SEP> Es gab keine Hinweise darauf, dass Lamb, der zwei Online-Kurse für die Delta State University in Cleveland, Mississippi, unterrichtete, bereits 300 Meilen zum Campus der Schule gereist war, wo er nach Polizeiangaben einen beliebten Geschichtsprofessor, Ethan Schmidt, in der Tür zu seinem Büro erschossen haben soll.<SEP> Der Polizeichef der Delta State University, Lynn Buford, sagte, die Beamten der Universität hätten von der Schießerei um 10.18 Uhr gehört.<SEP> Er sagte, Lamb habe den schicksalhaften 911-Anruf irgendwann danach gemacht.<SEP> Am Ende des Tages werde es einen weiteren Tod geben: Lamb habe sich das Leben genommen, als die Polizei ihn eingeschlossen habe.<SEP> Einen Tag nach der Schießerei in der Schule, die Studenten und Dozenten gezwungen hatte, sich hinter verschlossenen Türen zu verstecken, versuchten die Behörden noch immer zusammenzufügen, was Lamb motivierte.<SEP> Die von den Ermittlern an beiden Enden des Staates veröffentlichten Details sowie Studenten und Angestellte, die ihn kannten, halfen dabei, ein Bild eines talentierten, aber möglicherweise aufgewühlten Lehrers zu malen.<SEP> Studenten sagten, sie freuten sich auf seine Klasse.<SEP> Die Polizei in Gautier, wo Prentiss starb, sagte, er habe keine Geschichte von Gewalt oder Vorstrafen gehabt.<SEP> Schmidt selbst hatte Lamb in ein von ihm verfasstes Buch aufgenommen, in dem er die "wunderbaren Menschen" würdigte, mit denen er sein akademisches Leben teilte.<SEP> Beide lehrten in der Abteilung für Sozialwissenschaften und Geschichte, die 17 Fakultätsmitglieder auflistet, und viele Studenten nahmen Kurse von beiden.<SEP> Gleichzeitig gab es einige Neigungen zu Problemen.<SEP><BRK>

# English Translation – Document-level

In 911-Anruf gibt Professor zu, Freundin erschossen zu haben

In einem 911-Anruf sei seine Stimme nur leicht wackelig gewesen, sagte College-Professor Shannon Lamb der Polizei, er habe seine Freundin erschossen und Beamte hätten in ihr Haus eindringen müssen.

Lamb gab zu Protokoll, sein "süßer Hund" sei lebend da gewesen und vermutlich aufgebracht gewesen und sagte, die Familienkontakte der toten Frau seien auf ihrem Handy gefunden worden.

Im Inneren des Hauses fanden Beamte die Leiche von Amy Prentiss und einen handschriftlichen Zettel, der auf ein weißes Polster gekritzelt war: "Es tut mir so leid, dass ich wünschte, ich könnte sie zurücknehmen, ich liebte Amy und sie ist die einzige Frau, die mich jemals geliebt hat", hieß es in dem Brief, den die Behörden unterschrieben hatten.

Es gab keine Hinweise darauf, dass Lamb, der zwei Online-Kurse für die Delta State University in Cleveland, Mississippi, unterrichtete, bereits 300 Meilen zum Campus der Schule gereist war, wo er nach Polizeiangaben einen beliebten Geschichtsprofessor, Ethan Schmidt, in der Tür zu seinem Büro erschossen haben soll.

Der Polizeichef der Delta State University, Lynn Buford, sagte, die Beamten der Universität hätten von der Schießerei um 10.18 Uhr gehört.

Er sagte, Lamb habe den schicksalhaften 911-Anruf irgendwann danach gemacht.

Am Ende des Tages werde es einen weiteren Tod geben: Lamb habe sich das Leben genommen, als die Polizei ihn eingeschlossen habe.

| Model | Parameters | Layers | Dim |
|---|---|---|---|
| BERT/GPT-2 | 117M | 12 | 768/4096 |
| BERT/GPT-2 | 345M | 24 | 1024/4096 |
| GPT-2 | 762M | 36 | 1280/4096 |
| GPT-2 | 1542M | 48 | 1600/4096 |

| Model | Parameters | Layers | Dim |
|---|---|---|---|
| Nematus RNN | 25M (95MB) | 1/1 (2/2) | 512/1024 |
| Transformer (Base) | 30M (117MB) | 6/6 | 512/2048 |
| Transformer (Big) | 209M (790MB) | 6/6 | 1024/4096 |
| Transformer (Bigger) | 386M (1,471MB) | 12/12 | 1024/4096 |
| Transformer (Even Bigger) | 570M | 18/18 | 1024/4096 |
| Transformer (Biggest) | 750M | 24/24 | 1024/4096 |

# Scale down!

# Sequence-Level Knowledge Distillation

**Yoon Kim**
yoonkim@seas.harvard.edu

**Alexander M. Rush**
srush@seas.harvard.edu

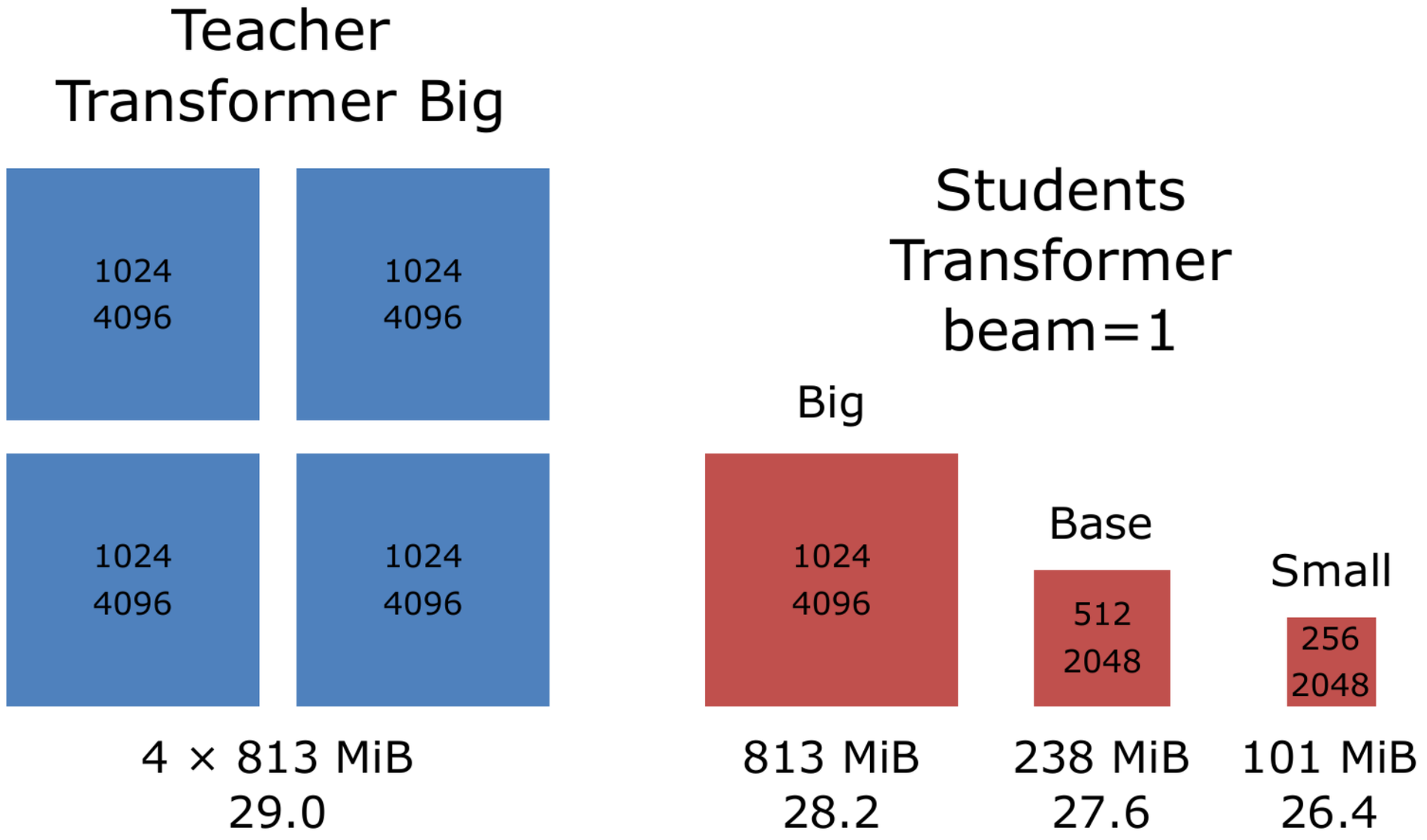
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA

# Marian: Cost-effective High-Quality Neural Machine Translation in C++

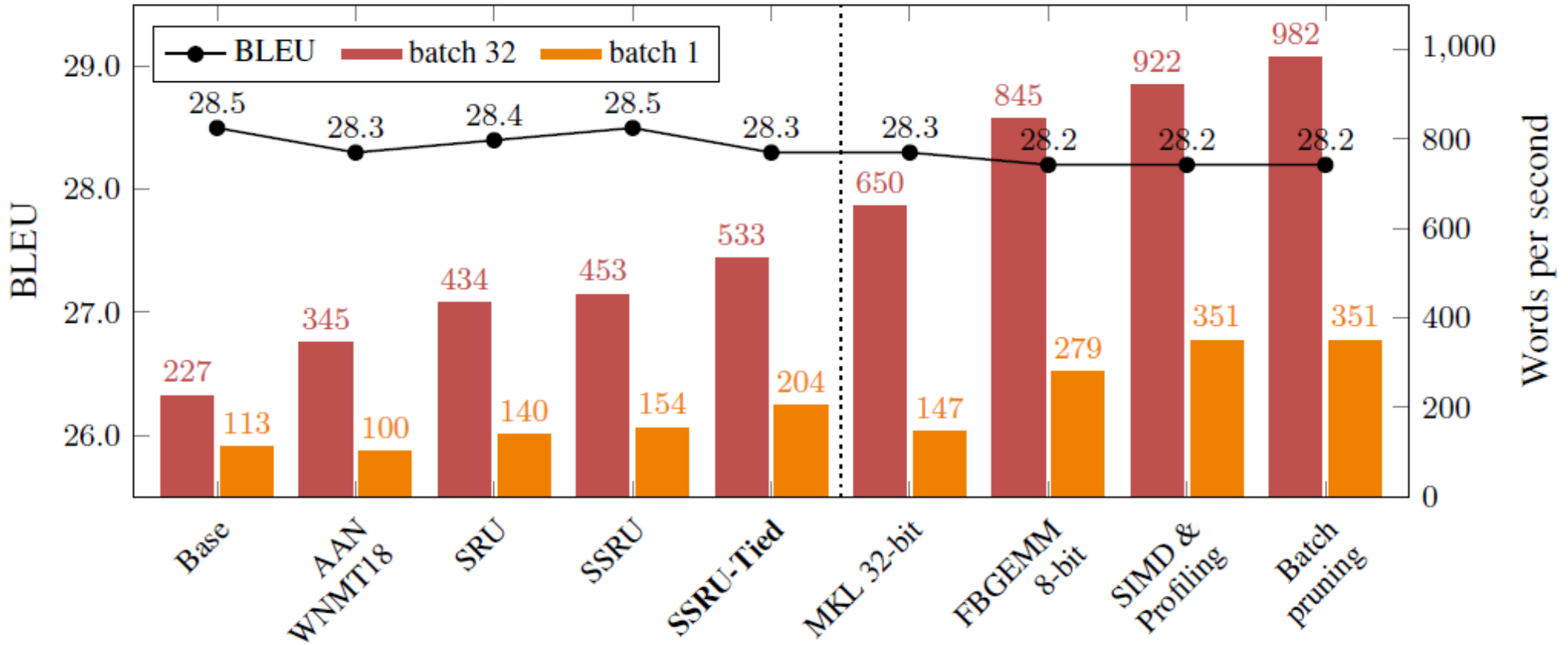
**Marcin Junczys-Dowmunt**[†] **Kenneth Heafield**[‡]
**Hieu Hoang**[‡] **Roman Grundkiewicz**[‡] **Anthony Aue**[†]

[†]Microsoft Translator
1 Microsoft Way
Redmond, WA 98121, USA

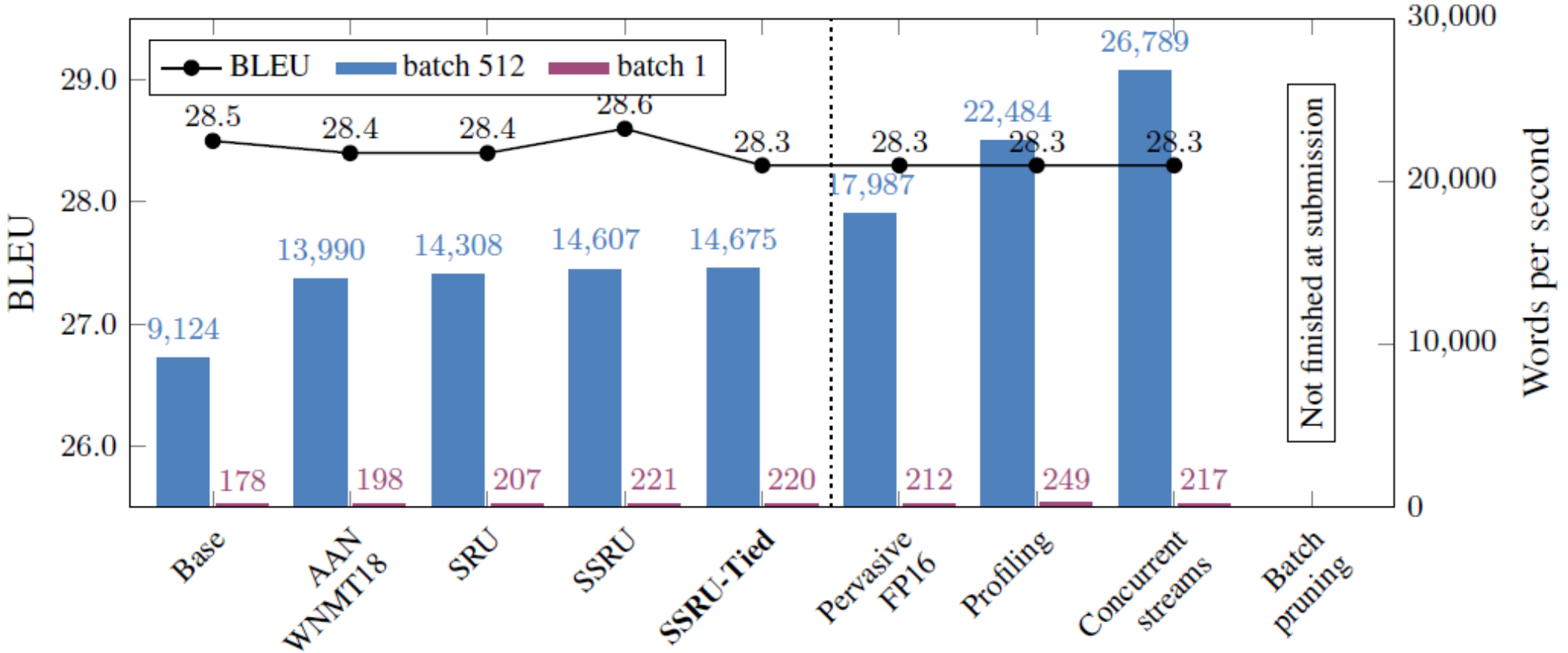
[‡]University of Edinburgh
10 Crichton Street
Edinburgh, Scotland, EU




(Scale preserving)




(a) Performance on a single CPU core and thread for newstest2014 on AWS m5.large, dedicated instance




(b) Performance on a NVidia Volta 100 GPU for newstest2014 on AWS p3.x2large

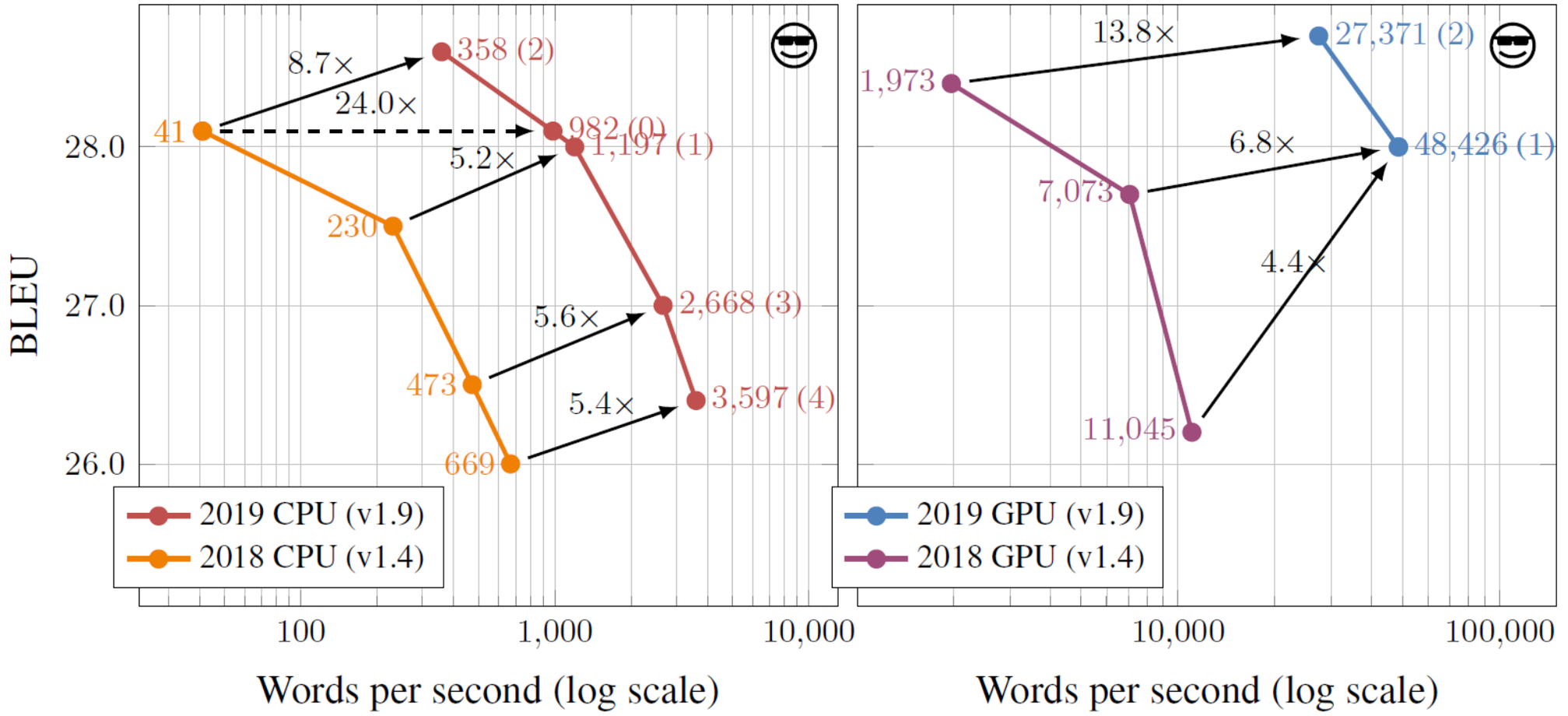



Figure 2: Relative speed improvements for fastest Marian models of comparable or better quality than submissions to WNMT2018 on newstest2014. Numbers in parentheses next to words-per-second values correspond to numbered submissions in Table 3. We also include our unsubmitted in-production model (0).

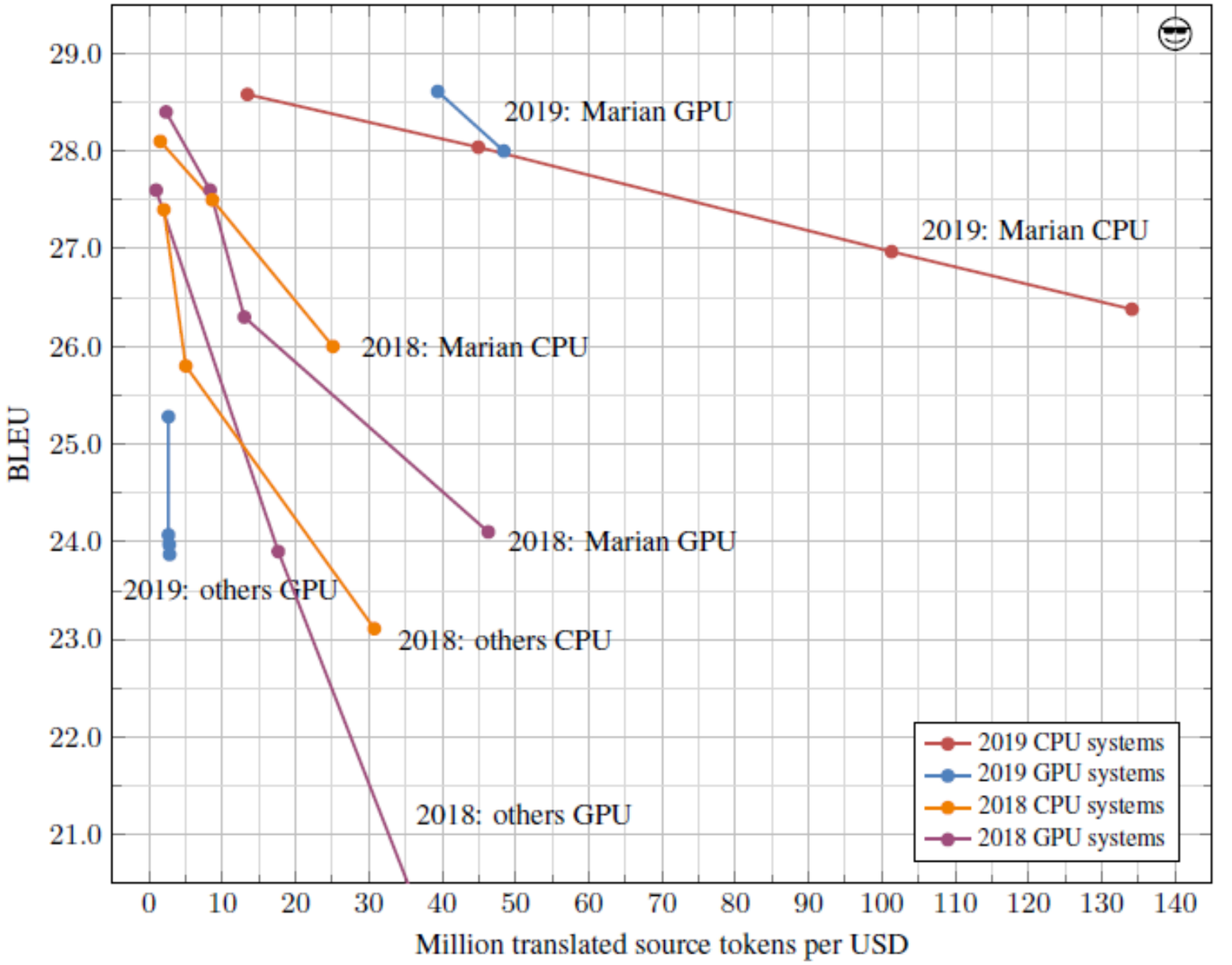



Figure 3: Pareto frontier for cost-effectiveness vs BLEU for all submissions (ours and other participants) from 2018 and 2019 on newstest2014 as reported by the organizers. We omit the weak baselines.

# Thank you!

Microsoft.com/Translator

blogs.msdn.com/translator

twitter.com/MSTranslator

facebook.com/BingTranslator