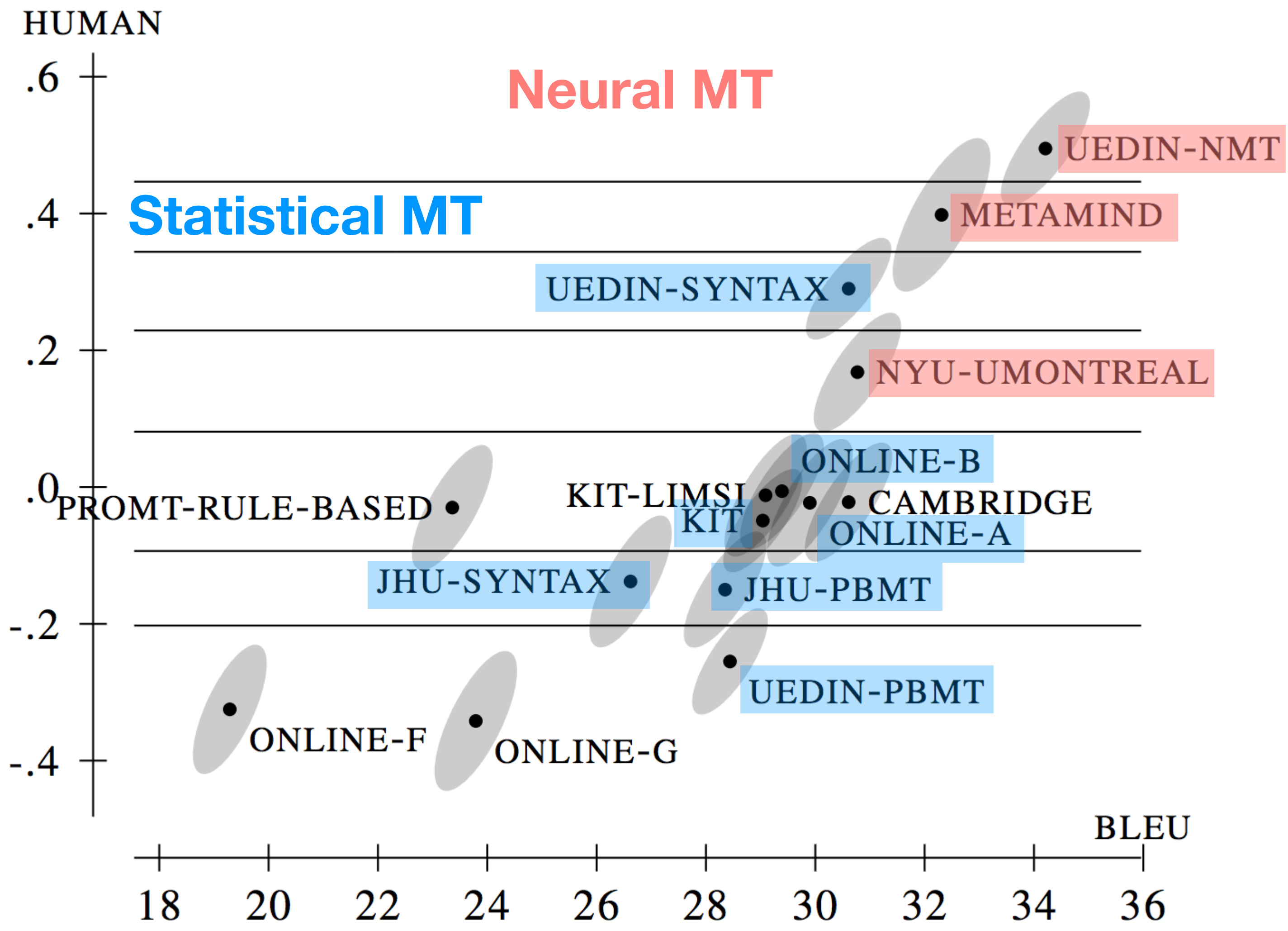
Current Challenges

Philipp Koehn

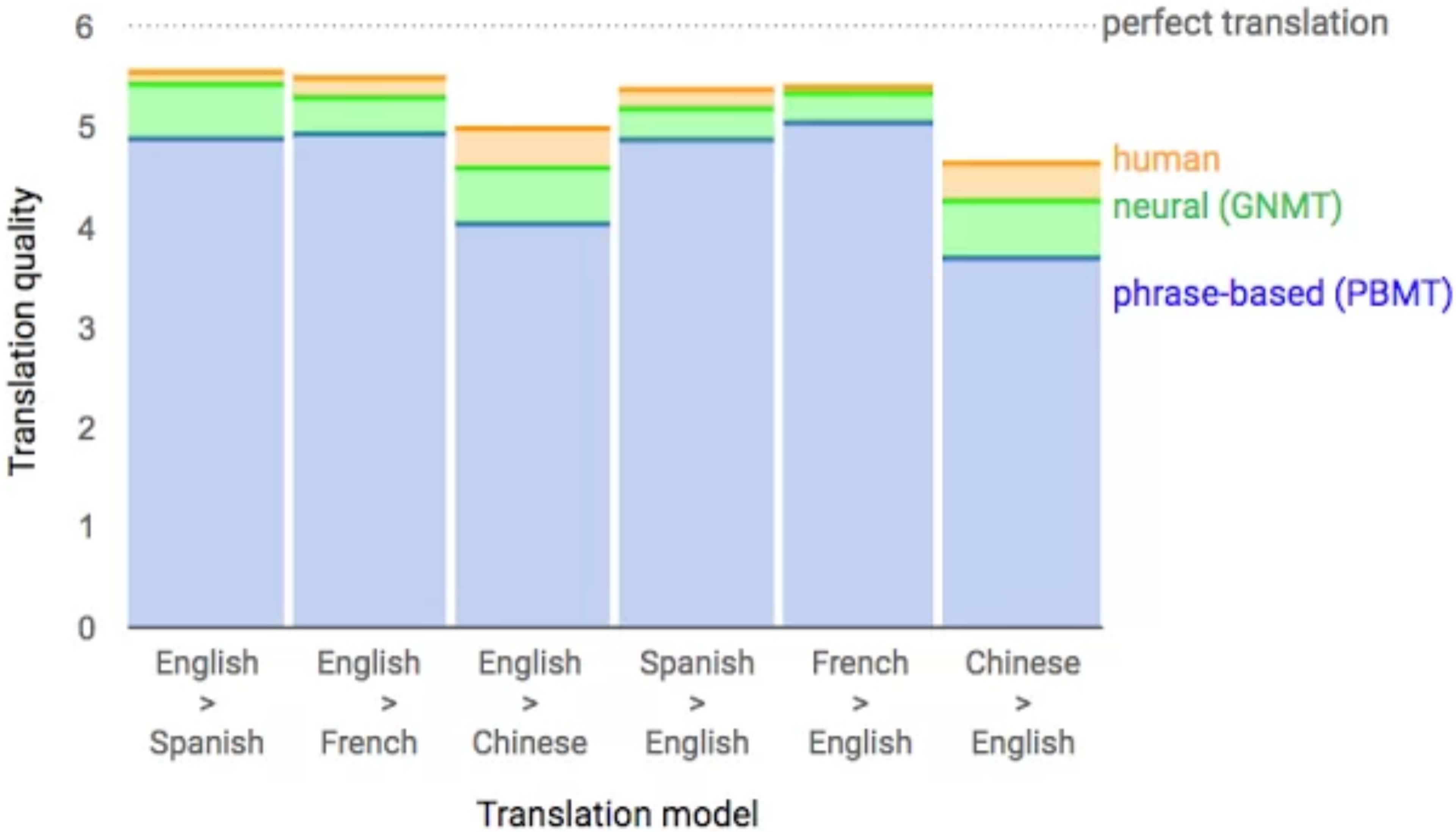
2 November 2023

WMT 2016



(in 2017 barely any statistical machine translation submissions)

2017: Google: "Near Human Quality"



2018: More Hype

Microsoft Research Achieves Human Parity For Chinese English Translation

Written by Sue Gee

Wednesday, 21 March 2018

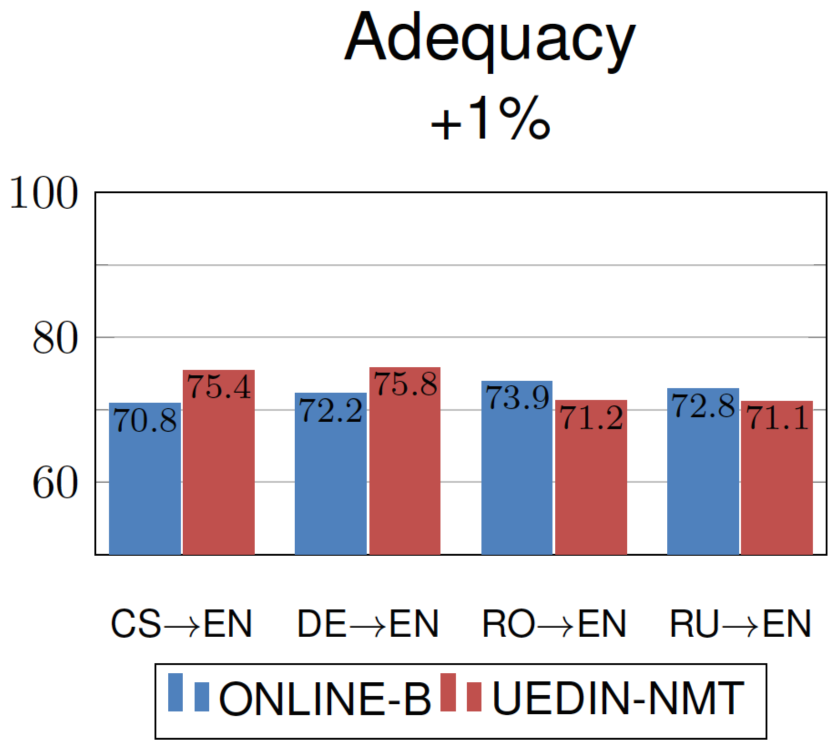
Researchers in Microsoft's labs in Beijing and in Redmond and Washington have developed an AI machine translation system that can translate with the same accuracy as a human from Chinese to English.

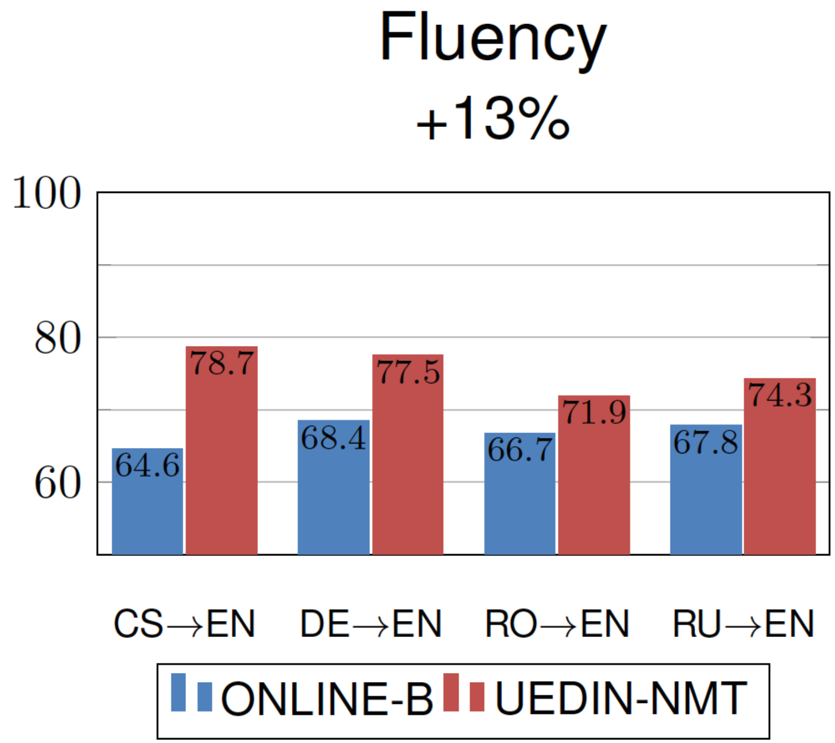
SDL Cracks Russian to English Neural Machine Translation

Global Enterprises to Capitalize on Near Perfect Russian to English Machine Translation as SDL Sets New Industry Standard

"90% of the system's output labelled as perfect by professional Russian-English translators"

Just Better Fluency?

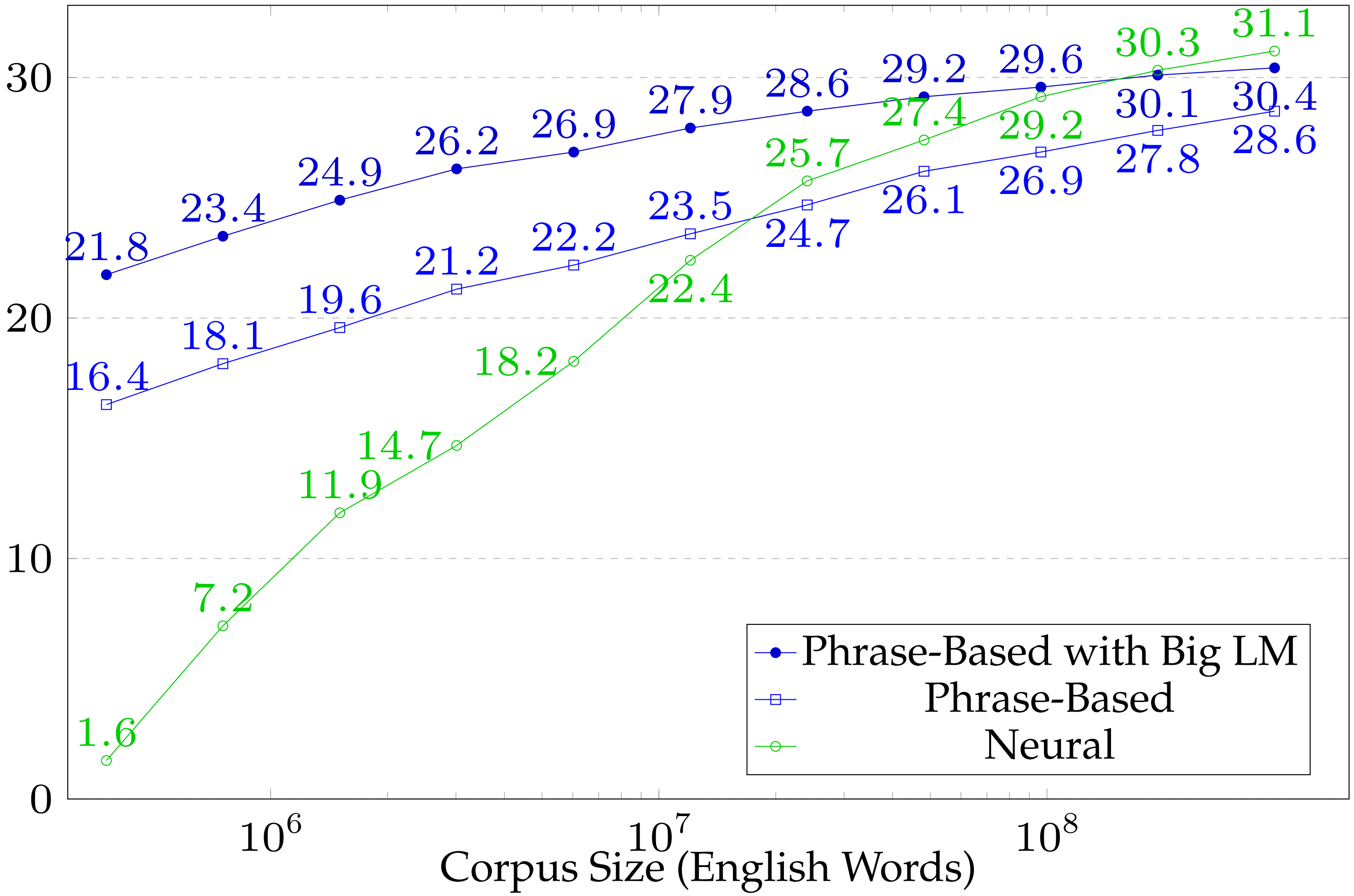




(from: Sennrich and Haddow, 2017)

lack of training data

Amount of Training Data



English-Spanish systems trained on 0.4 million to 385.7 million words

Translation Examples

Source	A Republican strategy to counter the re-election of Obama
$\frac{1}{1024}$	Un órgano de coordinación para el anuncio de libre determinación
$\frac{1}{512}$	Lista de una estrategia para luchar contra la elección de hojas de Ohio
$\frac{1}{256}$	Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor
$\frac{1}{128}$	Una estrategia republicana para la eliminación de la reelección de Obama
$\frac{1}{64}$	Estrategia siria para contrarrestar la reelección del Obama .
$\frac{1}{32}+$	Una estrategia republicana para contrarrestar la reelección de Obama

domain mismatch

Domain Mismatch

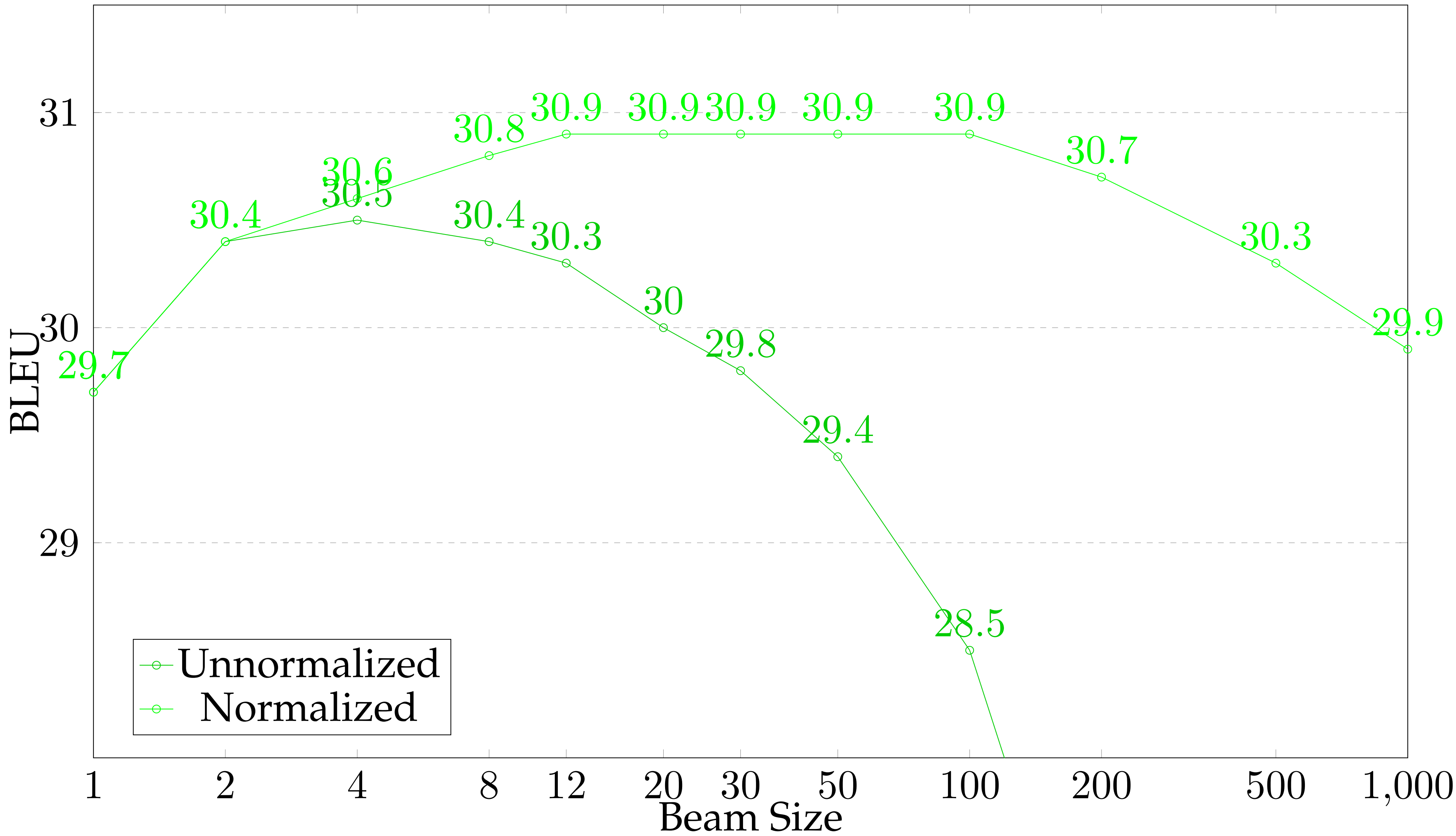
System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 32.8	45.1 42.2	35.3 44.7	17.9 17.9	26.4 20.8
Law	31.1 34.4	12.1 18.2	3.5 6.9	1.3 2.2	2.8 6.0
Medical	3.9 10.2	39.4 43.5	2.0 8.5	0.6 2.0	1.4 5.8
IT	1.9 3.7	6.5 5.3	42.1 39.8	1.8 1.6	3.9 4.7
Koran	0.4 1.8	0.0 2.1	0.0 2.3	15.9 18.8	1.0 5.5
Subtitles	7.0 9.9	9.3 17.8	9.2 13.6	9.0 8.4	25.9 22.1

Translation Examples

Source	Schaue um dich herum.
Ref.	Look around you.
All	NMT: Look around you. SMT: Look around you.
Law	NMT: Sughum gravecorn. SMT: In order to implement dich Schaue .
Medical	NMT: EMEA / MB / 049 / 01-EN-Final Work progamme for 2002 SMT: Schaue by dich around .
IT	NMT: Switches to paused. SMT: To Schaue by itself . \t \t
Koran	NMT: Take heed of your own souls. SMT: And you see.
Subtitles	NMT: Look around you. SMT: Look around you .

beam search

Beam Search


BLEU
31
30
29
Beam Size
1
2
4
8
12
20
30
50
100
200
500
1,000
Unnormalized
Normalized
29.7
30.4
30.6
30.5
30.8
30.4
30.9
30.3
30.9
30
30.9
29.8
30.9
29.4
30.9
28.5
30.7
30.3
29.9

noisy data

Noise in Training Data

- Crawled parallel data from the web (very noisy)

	SMT	NMT
WMT17	24.0	27.2
+ Paracrawl	25.2 (+1.2)	17.3 (-9.9)

(German-English, 90m words each of WMT17 and Crawl data)

	5%		10%		20%		50%		100%	
Raw crawl data	27.4	24.2	26.6	24.2	24.7	24.4	20.9	24.8	17.3	25.2
	+0.2	+0.2	-0.9	+0.2	-2.5	+0.4	-6.3	+0.8	-9.9	+1.2

- Corpus cleaning methods [Xu and Koehn, EMNLP 2017] give improvements

Types of Noise

- Misaligned sentences
- Disfluent language (from MT, bad translations)
- Wrong language data (e.g., French in German–English corpus)
- Untranslated sentences
- Short segments (e.g., dictionaries)
- Mismatched domain

Mismatched Sentences

- Artificial created by randomly shuffling sentence order
- Added to existing parallel corpus in different amounts

5%		10%		20%		50%		100%	
	24.0		24.0		23.9	26.1	23.9	25.3	23.4
	-0.0		-0.0		-0.1	-1.1	-0.1	-1.9	-0.6

- Bigger impact on NMT (green, left) than SMT (blue, right)

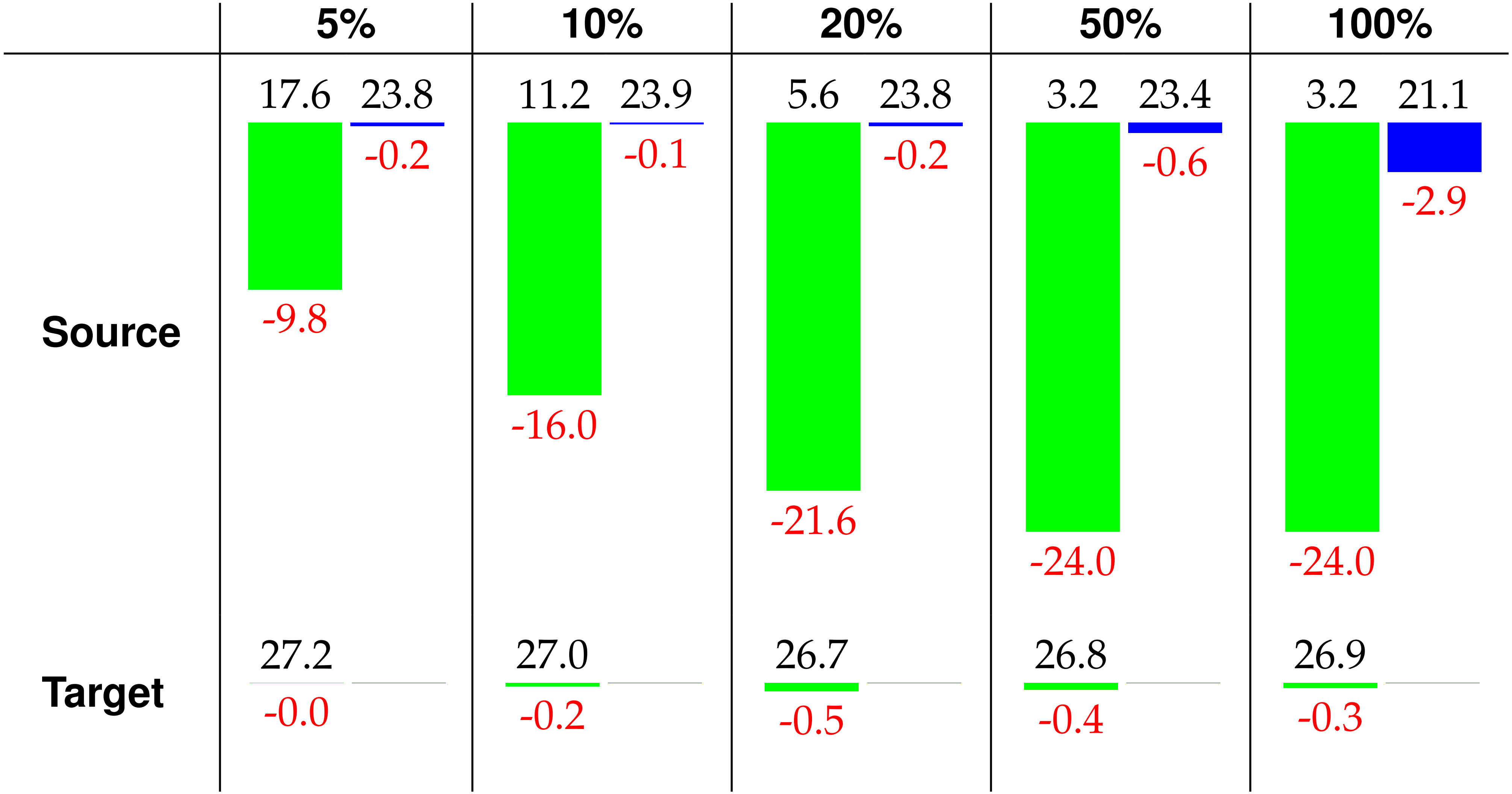
Misordered Words

- Artificial created by randomly shuffling words in each sentence

	5%		10%		20%		50%		100%	
Source		24.0 -0.0		23.6 -0.4		23.9 -0.1	26.6 -0.6	23.6 -0.4	25.5 -1.7	23.7 -0.3
Target		24.0 -0.0		24.0 -0.0		23.4 -0.6	26.7 -0.5	23.2 -0.8	26.1 -1.1	22.9 -1.1

- Similar impact on NMT than SMT, worse for source reshuffle

Untranslated Sentences



Wrong Language

	5%		10%		20%		50%		100%	
fr source	26.9 -0.3	24.0 -0.0	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.9 -0.1	26.8 -0.4	23.8 -0.2
fr target	26.7 -0.5	24.0 -0.0	26.6 -0.6	23.9 -0.1	26.7 -0.5	23.8 -0.2	26.2 -1.0	23.5 -0.5	25.0 -2.2	23.4 -0.6

- Surprisingly robust, maybe due to domain mismatch of French data

Short Sentences

	5%		10%		20%		50%	
1-2 words	27.1 -0.1	24.1 +0.1	26.5 -0.7	23.9 -0.1	26.7 -0.5	23.8 -0.2		
1-5 words	27.8 +0.6	24.2 +0.2	27.6 +0.4	24.5 +0.5	28.0 +0.8	24.5 +0.5	26.6 -0.6	24.2 +0.2

- No harm done

control over output

Specifying Decoding Constraints

- Overriding the decisions of the decoder
- Why?

  $\Rightarrow$ translations have followed strict terminology

  $\Rightarrow$ rule-based translation of dates, quantities, etc.

XML Schema

The <x translation="Router"> router </x> is <wall/>

a model <zone> Psy X500 Pro </zone> .

- The XML tags specify to the decoder that
  - the word router to be translated as Router
  - The router is, to be translated before the rest (<wall/>)
  - brand name Psy X500 Pro to be translated as a unit (<zone>, </zone>)

Formal Constraints

- Subtitles
  - translation has to fit into space on screen (may have to be shortened)
  - input and output broken up into lines
- Speech translation
  - input often not well-formed
  - real time translation: start while sentence is spoken
  - subtitles: have be readable in limited time
  - dubbing: sync up with video of speaker's mouth movement
- Poetry
  - meter
  - rhyme

catastrophic errors

Catastrophic Errors

News | Science and Technology

Facebook apologises for rude mistranslation of Xi Jinping's name

Company blames technical glitch that 'caused incorrect translations' of Chinese leader's name from Burmese to English.

Facebook's auto translation AI fail leads to a nightmare for a Palestinian man

The AI feature had "Good morning" in Arabic wrongly translated as "attack them" in Hebrew.

By Gianluca Mezzofiore on October 24, 2017

Industry News • By Marion Marking On 3 Aug 2020

Thai Mistranslation Shows Risk of Auto-Translating Social Media Content

After a machine translation of a post from English into Thai about the King's birthday proved offensive to the Thai monarchy, Facebook Thailand said it was deactivating auto-translate on Facebook and Instagram, revamping machine translation (MT) quality, and offering the Thai people its "profound apology."

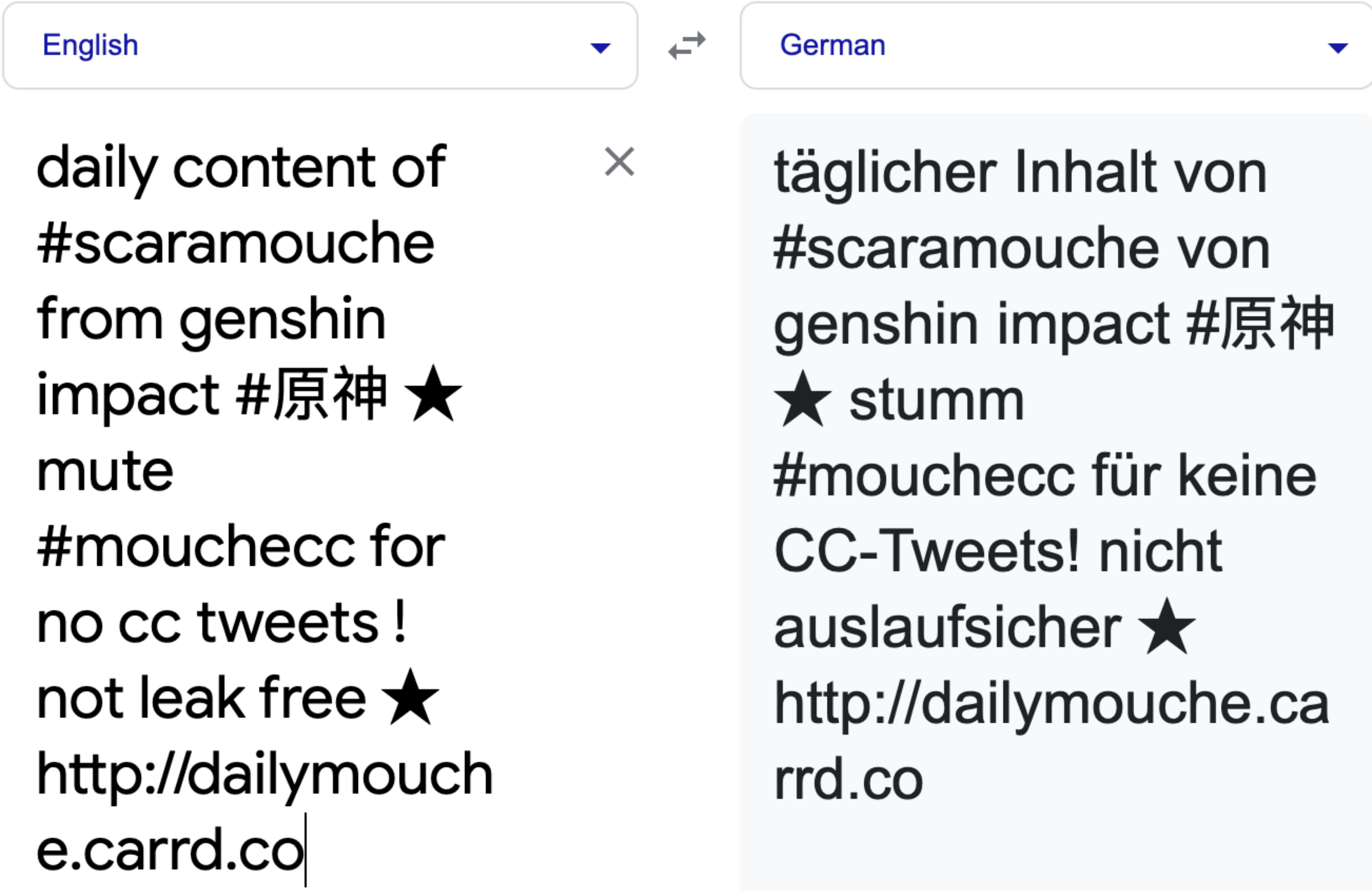
What are Catastrophic Errors?

- Generation of profanity
  - first step: maintain list of offensive words for each language
  - only eliminate these words, if the input did not include such words
  - but: offensive language is not limited to specific words
- Generation of violent / inciting content
- Opposite meaning
- Mistranslation of names

$\Rightarrow$ All this is hard to detect

robustness

Robustness to User Generated Content



Challenges

- Jargon and acronyms
- Misspellings (sometimes intended for effect)
- Mangled grammar
- Special symbols (emojis, etc.)
- Hashtags, URLs, ...
- Use of dialectical languages
- Use of non-standard writing systems (e.g., Latin script due to lack of keyboard)

Some Methods

- Special handling of non-words like emojis, hashtags, URLs
- Creating synthetic noisy training data
- Adversarial training
- Resources
  - Machine translation of noisy text data set (MTNT)
  - WMT 2020 Shared Task on Machine Translation Robustness

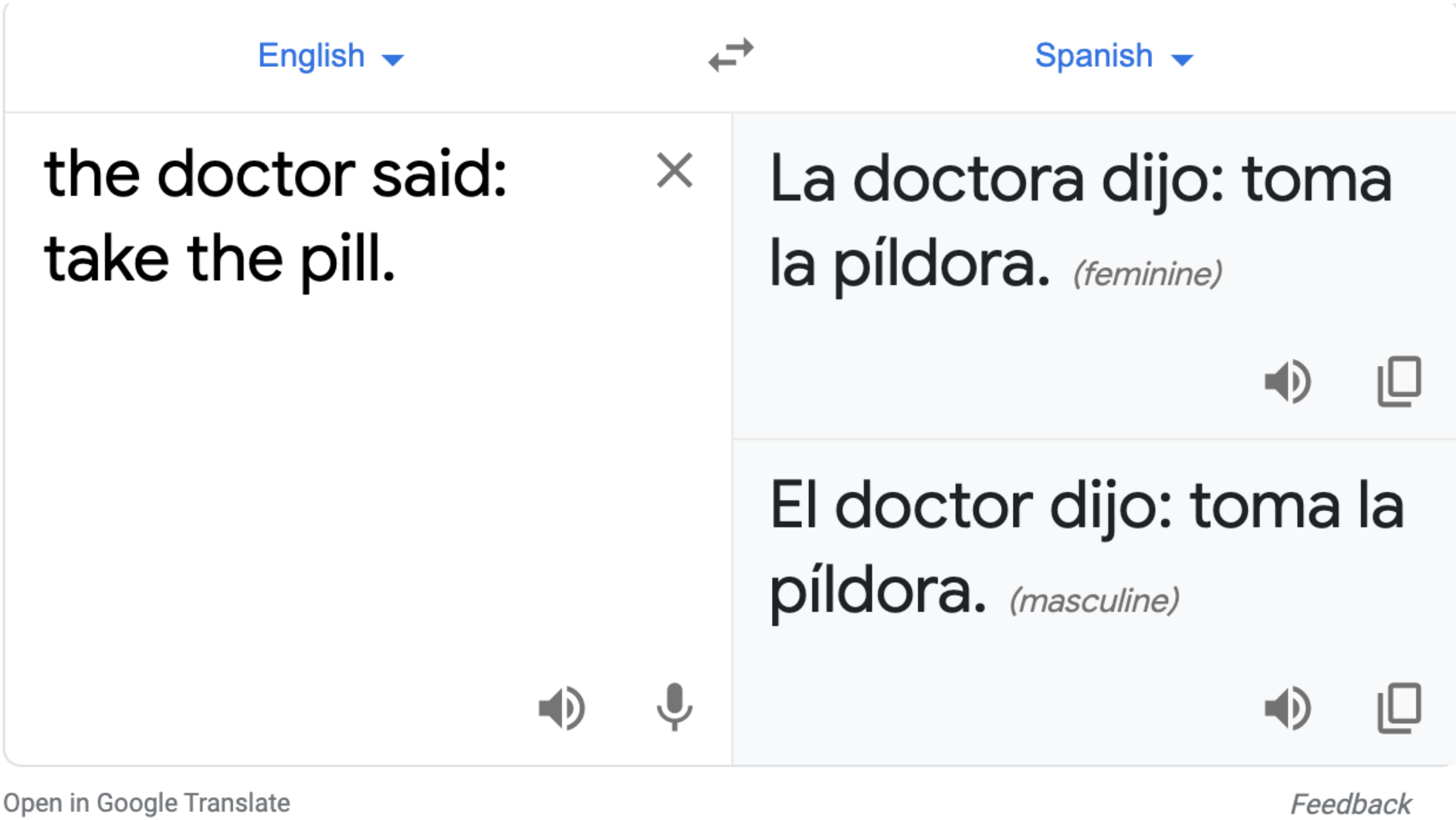
bias

Gender Bias

The doctor asked the nurse to help her in the procedure

El doctor le pidio a la enfermera que le ayudara con el procedimiento

Gender Bias

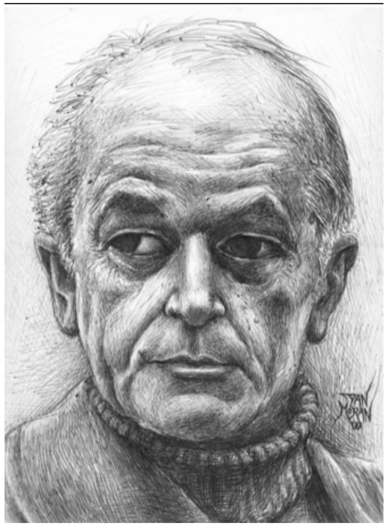

English
Spanish
the doctor said: take the pill.
La doctora dijo: toma la píldora. (feminine)
El doctor dijo: toma la píldora. (masculine)
Open in Google Translate
Feedback

Robustness to Style

"You Sound Just Like Your Father"
Commercial Machine Translation Systems Include Stylistic Biases

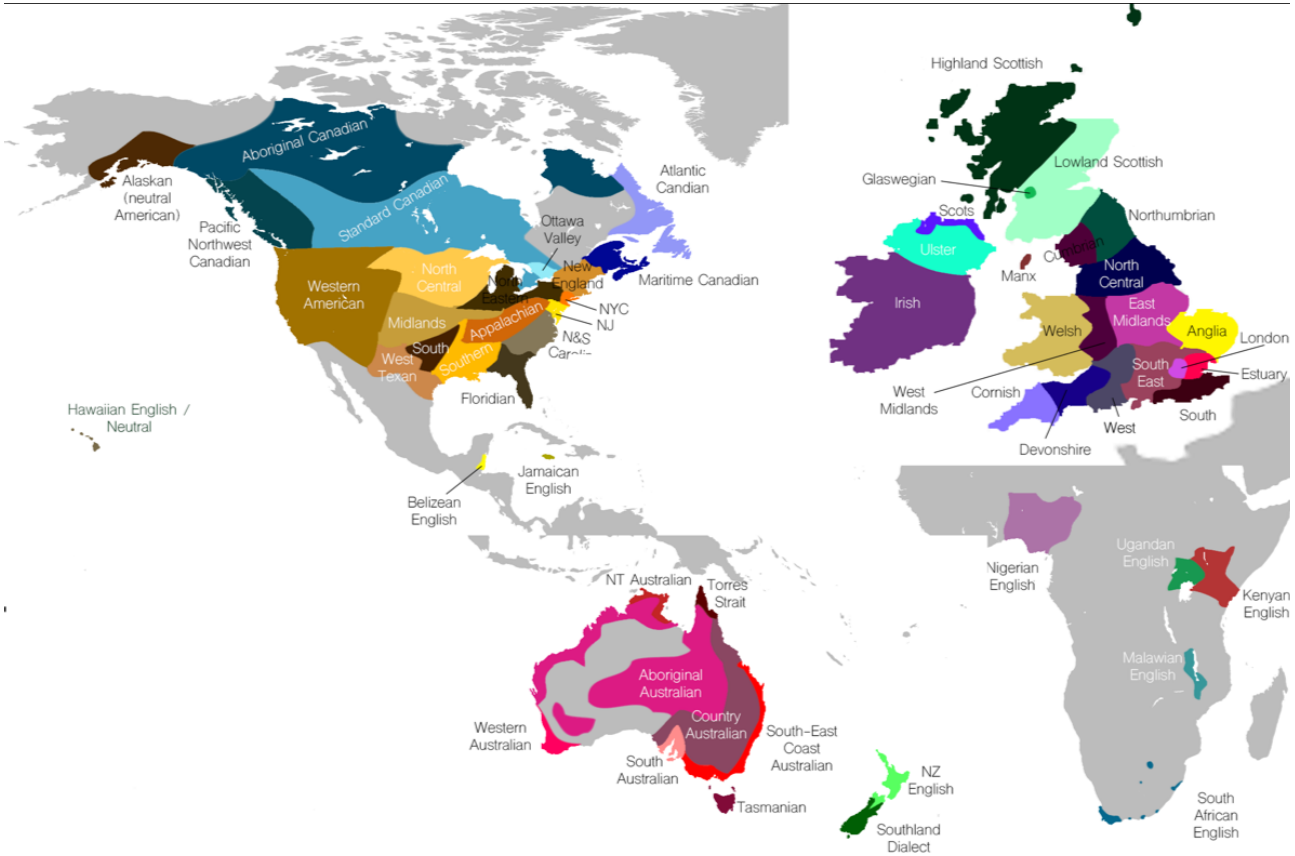
Dirk Hovy **Federico Bianchi** **Tommaso Fornaciari**
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
{dirk.hovy, f.bianchi, fornaciari.tommaso}@unibocconi.it

$\Rightarrow$



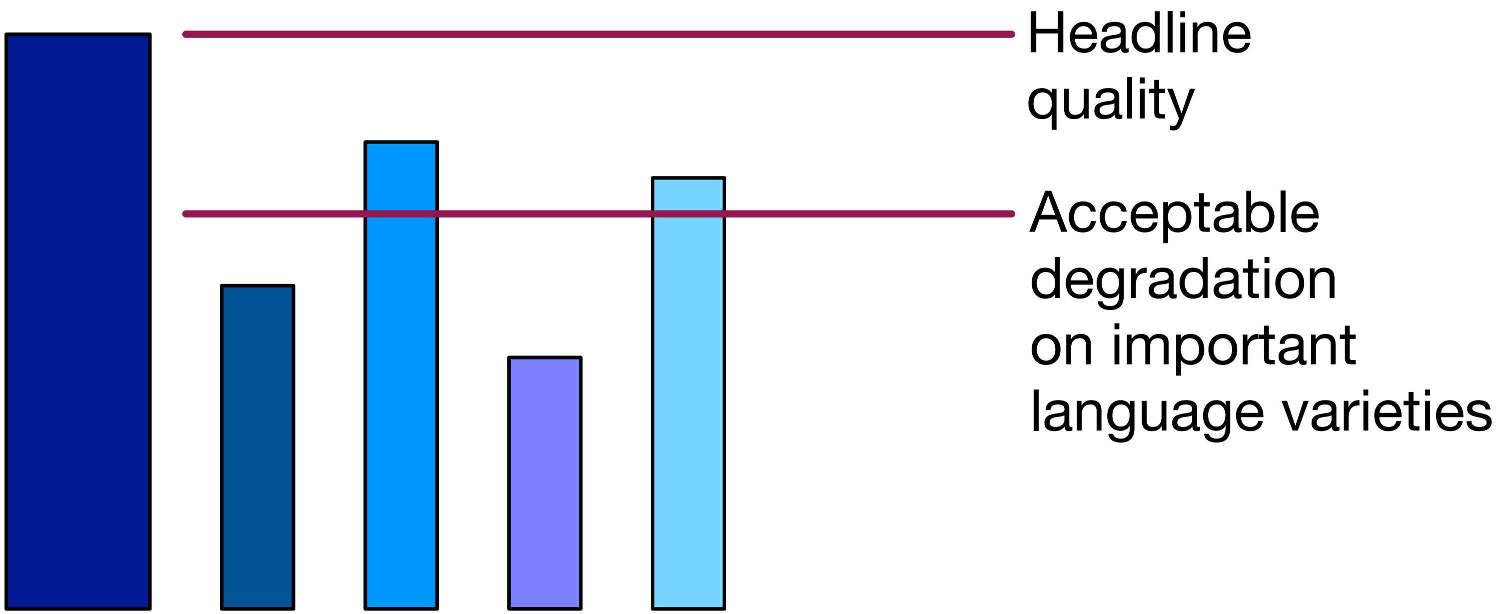
Dialect Bias

- Models often trained only on standard languages (British, American)
- Work less well on other dialects
- Bigger problem for automatic speech recognition



Evaluate Across Language Varieties

- BLEU score on standard language is not enough
- Also need test sets for each language variety



document-level translation

Document-Level Translation

The shop is selling a nice table.
Jane is quite taken by it.
The table would match the chairs in her living room.

- Machine translation translates one sentence at a time
- But: surrounding context may help

Document-Level Translation

The shop is selling a nice table.
Jane is quite taken by it.
The table would match the chairs in her living room.

- Machine translation translates one sentence at a time
- But: surrounding context may help
  - translation of pronouns may require co-reference

Document-Level Translation

The shop is selling a nice table.
Jane is quite taken by it.
The table would match the chairs in her living room.

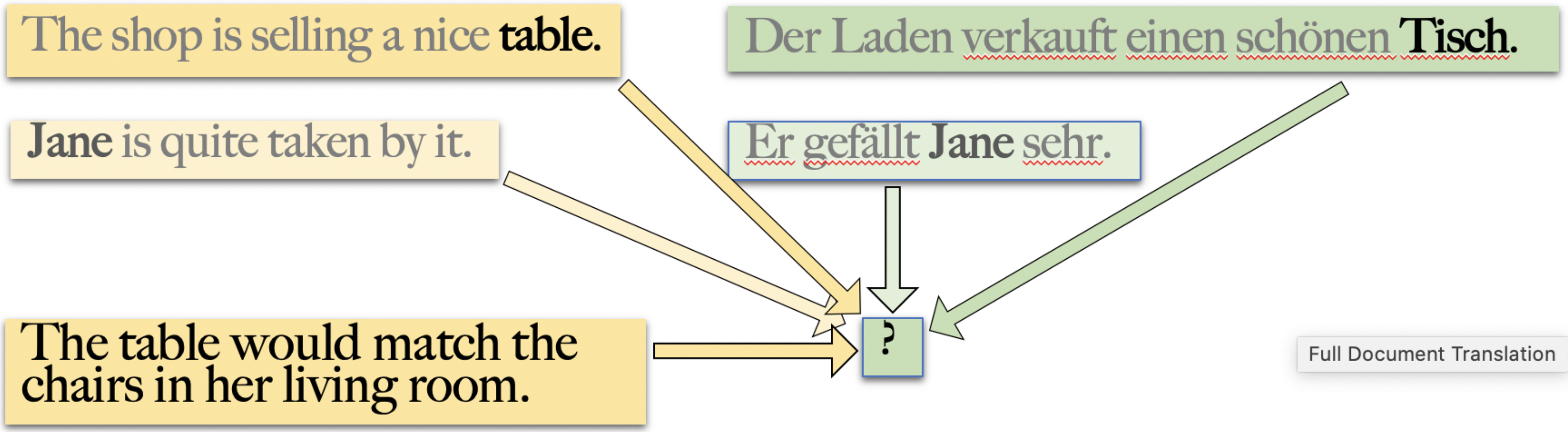
- Machine translation translates one sentence at a time
- But: surrounding context may help
  - translation of pronouns may require co-reference
  - ambiguous words may be informed by broader context

Document-Level Translation

The shop is selling a nice table.
Jane is quite taken by it.
The table would match the chairs in her living room.

- Machine translation translates one sentence at a time
- But: surrounding context may help
  - translation of pronouns may require co-reference
  - ambiguous words may be informed by broader context
  - consistent translation of repeated words

Conditioning on Broader Context



- Hierarchical attention
  - compute which previous sentences matter most
  - compute which words in these sentences matter most

Conditioning on Broader Context

The shop is selling a nice table. <s> Jane is quite taken by it. <s> The table would match the chairs in her living room.

↓

Der Laden verkauft einen schönen Tisch. <s> Er gefällt Jane sehr. <s> ...

- Concatenate all sentences together
  - document = very long sentence
  - special treatment for sentence boundaries
  - requires scaling of neural decoding implementation

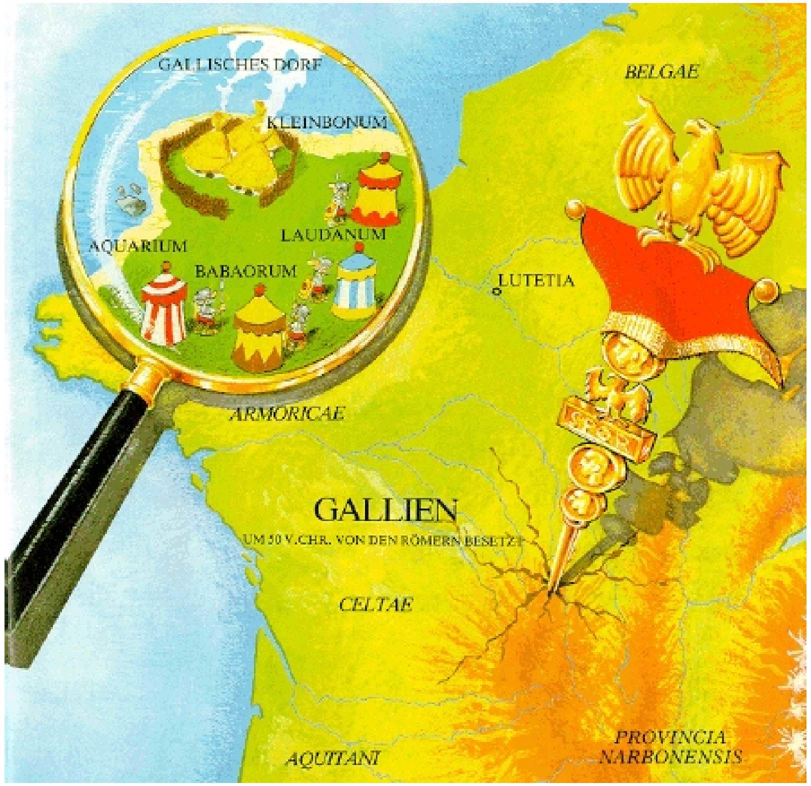
machine translation

and

large language models

The Large Language Model Wave

- Large language models have overtaken much of NLP
- So far, Machine Translation is still a hold-out: dedicated models are trained from scratch
- How long will this still be the case?



LMs as Unsupervised Learners (2018)

Language Models are Unsupervised Multitask Learners

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

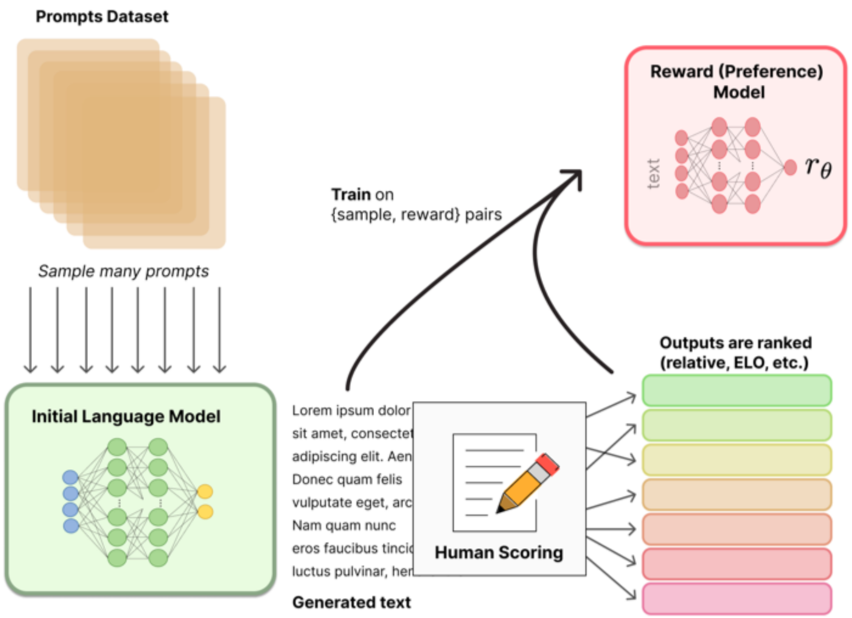
- Train language models on relatively clean text data (GPT-2)
- Convert any NLP problem into a text continuation problem
  - pre prompt engineering
  - goes into some detail of how each task is converted
  - impressive performance on many tasks
- Terrible at translation
  ... but all non-English text was removed from training corpus

Three Stages of Training Large Language Model

- Stage 1: Train on massive amounts of text (up to a trillion words)

- Stage 2: Instruction training
  - Examples of requests / responses constructed by human annotators
  - *"Summarize the following: ..."*
  - *"Give me ten examples of ..."*
  - *"Translate from French into English: ..."*

- Stage 3: Reinforcement learning from human feedback
  - Machine generates multiple responses to a prompt
  - Human annotators rank them
  - Train a reward model from
  - Fine-tune model with reward model



A Closer Look at PaLM for MT (2022)

Prompting PaLM for Translation: Assessing Strategies and Performance

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, George Foster
Google Research
{vilar, freitag, colincherry, jmluo, vratnakar, fosterg}@google.com

- Exploration of examples used for prompting
- Evaluation with BLEU / BLEURT / MQM (human eval)
- WMT 2021 test set for de,zh→en, WMT 2014 for fr→en

Examples Used for Prompting

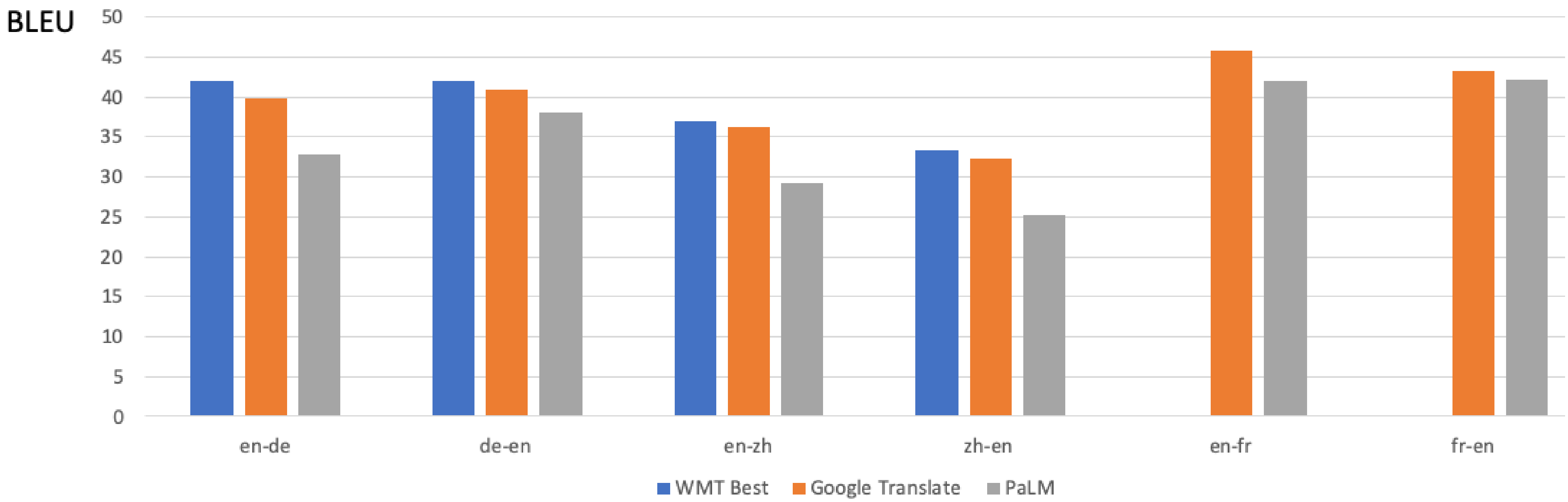
- Select parallel sentences from the WMT training data (full) or prior test sets (dev)

  random randomly pick sentence pairs
  kNN BOW prefer sentences pairs with lexical overlap on source
  kNN ROBERTa prefer pairs with similar ROBERTa embedding for source

LP	Pool	Selection	BLEURT	BLEU
en → de	full	random	71.8	32.9
		kNN BOW	71.7	32.4
		kNN RoBERTa	73.0	32.5
	dev	random	74.8	32.8
		kNN RoBERTa	74.8	32.3
de → en	full	random	74.8	38.4
		kNN BOW	72.7	36.9
		kNN RoBERTa	73.8	35.4
	dev	random	75.9	38.0
		kNN RoBERTa	75.8	37.2

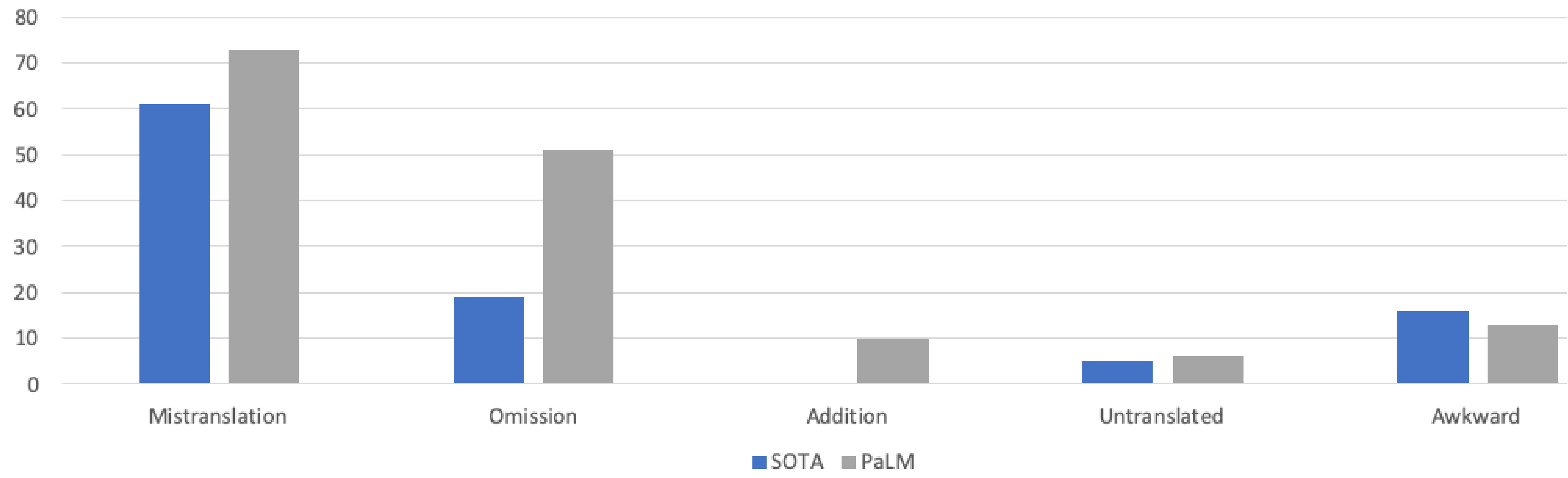
- BLEU likes full/random, BLEURT mixed bag

Comparison to State of the Art



Human Evaluation: MQM

- Language Models makes more adequacy errors, similar fluency
- German-English, MQM error categories (count of errors)



Translation Data in Training? (2023)

Searching for Needles in a Haystack:
On the Role of Incidental Bilingualism in PaLM's Translation Capability

Eleftheria Briakou ebriakou@cs.umd.edu
Colin Cherry colincherry@google.com
George Foster fosterg@google.com

- PaLM is exposed to over 30 million translation pairs across at least 44 languages
  - 1.4% of training examples are bilingual
  - 0.34% have a translated sentence pair
- Most bilingual content is code-switched, about 20% contains translations

Impact of Translation Data

- Sentence pairs can be extracted from bilingual samples
  - split sample into sentences
  - align English and French sentences with cross-lingual sentence embedding

  ⇒ parallel training corpus

- Training on mined parallel data (WMT fr-en): 38.1 BLEU
  Training on WMT training data: 42.0 BLEU

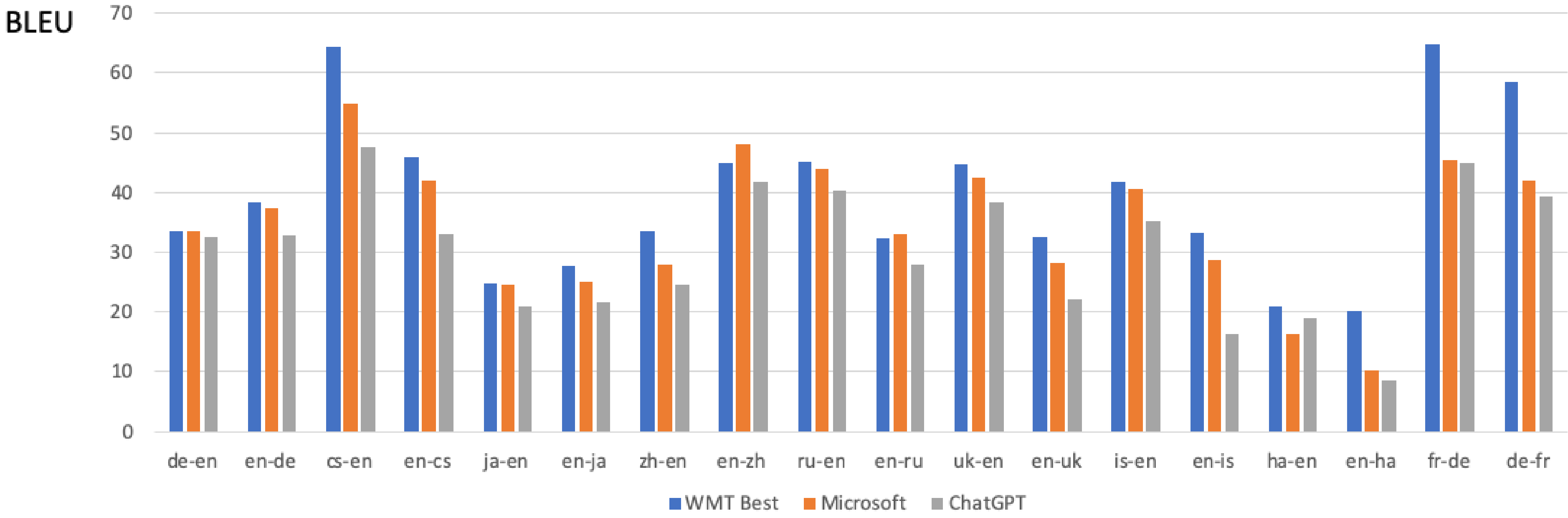
- Worse translation quality if bilingual content is removed from PaLM training

- Much worse translation quality with smaller (1B, 8B) PaLM models

How About GPT? (2023)

How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, Hany Hassan Awadalla*
Microsoft



One Strength: Document-Level Translation

- When translating multiple sentences at once, quality improves

System	COMET-22	COMETkiwi	Doc-COMETkiwi	ChrF	BLEU	Doc-BLEU	GPT Requests
			DE-EN				
WMT-Best	85.0	**81.4**	79.9	**58.5**	33.4	**35.2**	–
MS-Translator	84.7	81.0	79.5	**58.5**	**33.5**	**35.2**	–
GPT Sent ZS	84.8	81.2	79.5	56.8	30.9	32.3	1984
GPT Doc ZS w=2	85.1	**81.4***	80.0	57.8	32.6	34.4	1055
GPT Doc ZS w=4	**85.2***	81.3	**80.2***	57.9	32.8	34.5	607
GPT Doc ZS w=8	85.1	81.2	80.2	57.9	33.0	34.7	401
GPT Doc ZS w=16	85.2	81.2	80.2	58.0*	33.1*	34.8*	310
GPT Doc ZS w=32	85.1	81.2	80.2	57.9	33.1	34.8	274
			EN-DE				
WMT-Best	**87.2**	**83.6**	**83.1**	**64.6**	**38.4**	**40**	–
MS-Translator	86.8	83.4	83	64.2	37.3	38.8	–
GPT Sent ZS	85.6	82.8	82.2	60.2	31.8	33.1	2037
GPT Doc ZS w=2	86.1	82.7	82.4	60.9	32.8	34.4	1058
GPT Doc ZS w=4	86.3	82.6	82.6	61.3	33.6	35.2	579
GPT Doc ZS w=8	86.4	82.6	82.6	60.9	33.4	35.2	349
GPT Doc ZS w=16	86.5*	82.6*	82.6*	61.3*	34.2*	36.1*	235
GPT Doc ZS w=32	86.4	82.6	82.7	61.3	34.1	36.1	187

Convergence of LM and MT

- Both Language Models and Machine Translation are built with the same Transformer architecture

TRANSLATION	LANGUAGE
Der braune Hund is freundlich	*The [MASK] dog is [MASK].*
The brown dog is friendly	The brown dog is friendly

- This data can be mixed in any way

- Practical considerations: Large Language Models may be too big for use

questions?