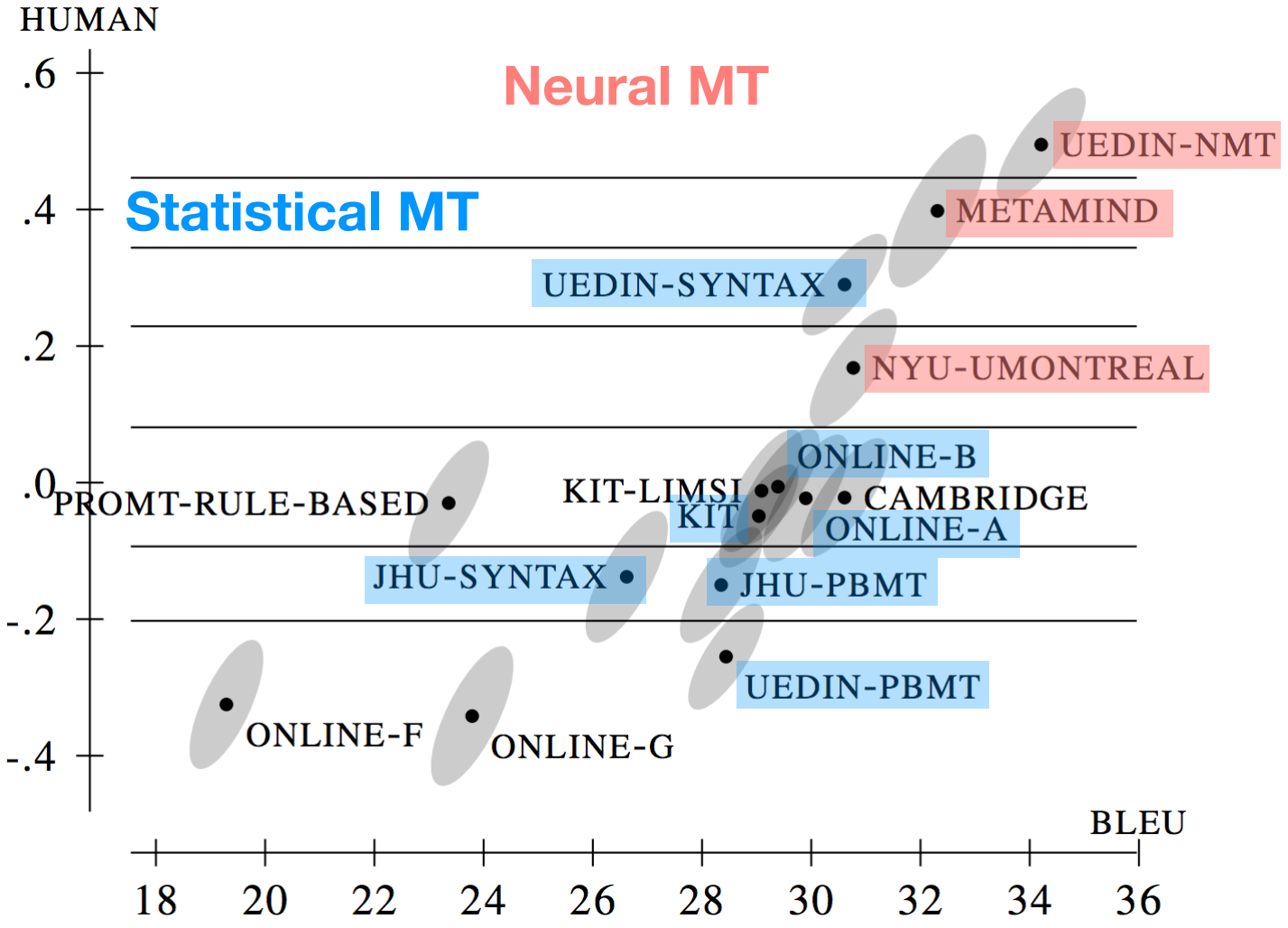
Current Challenges

Philipp Koehn

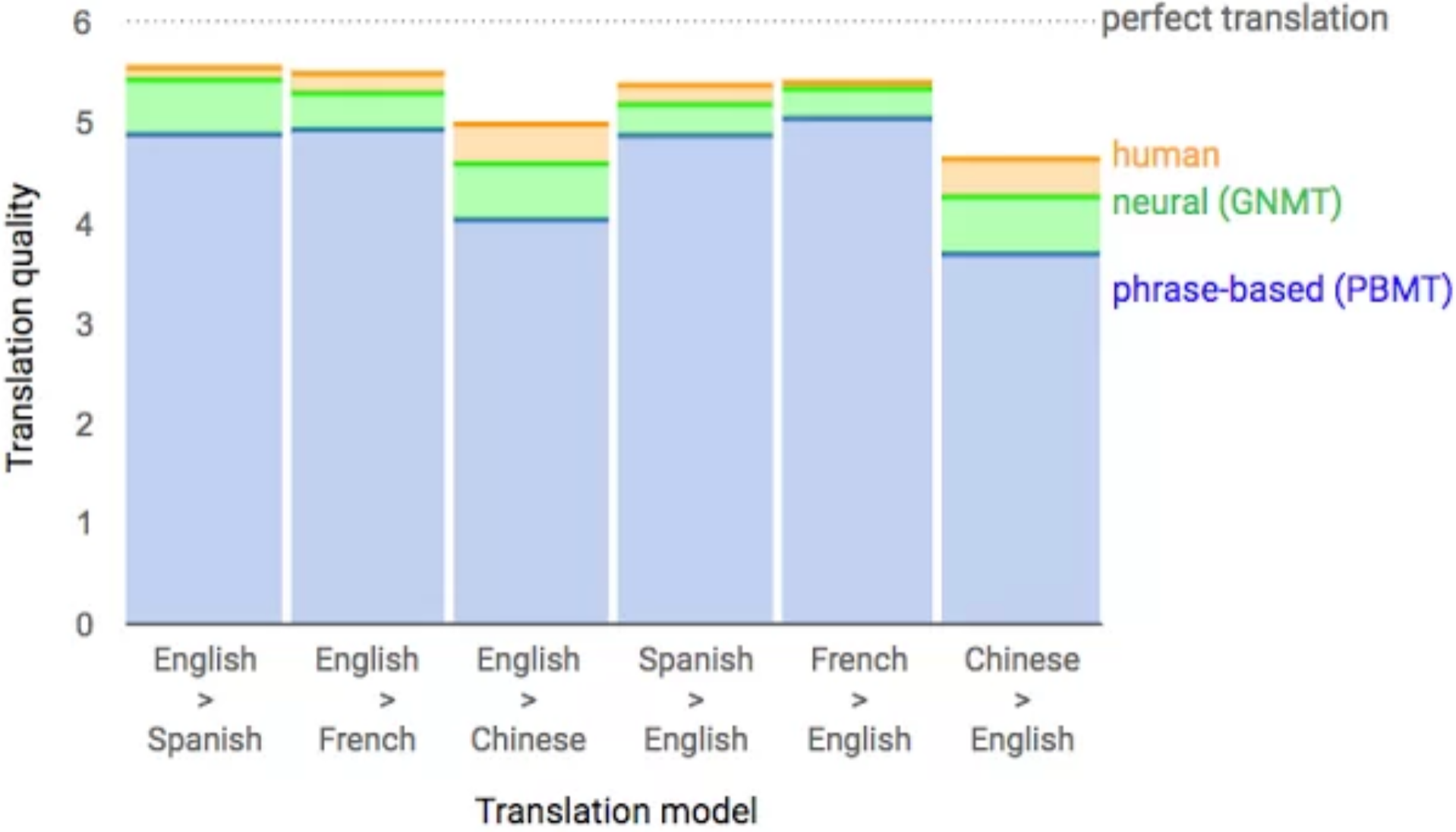
31 October 2024

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

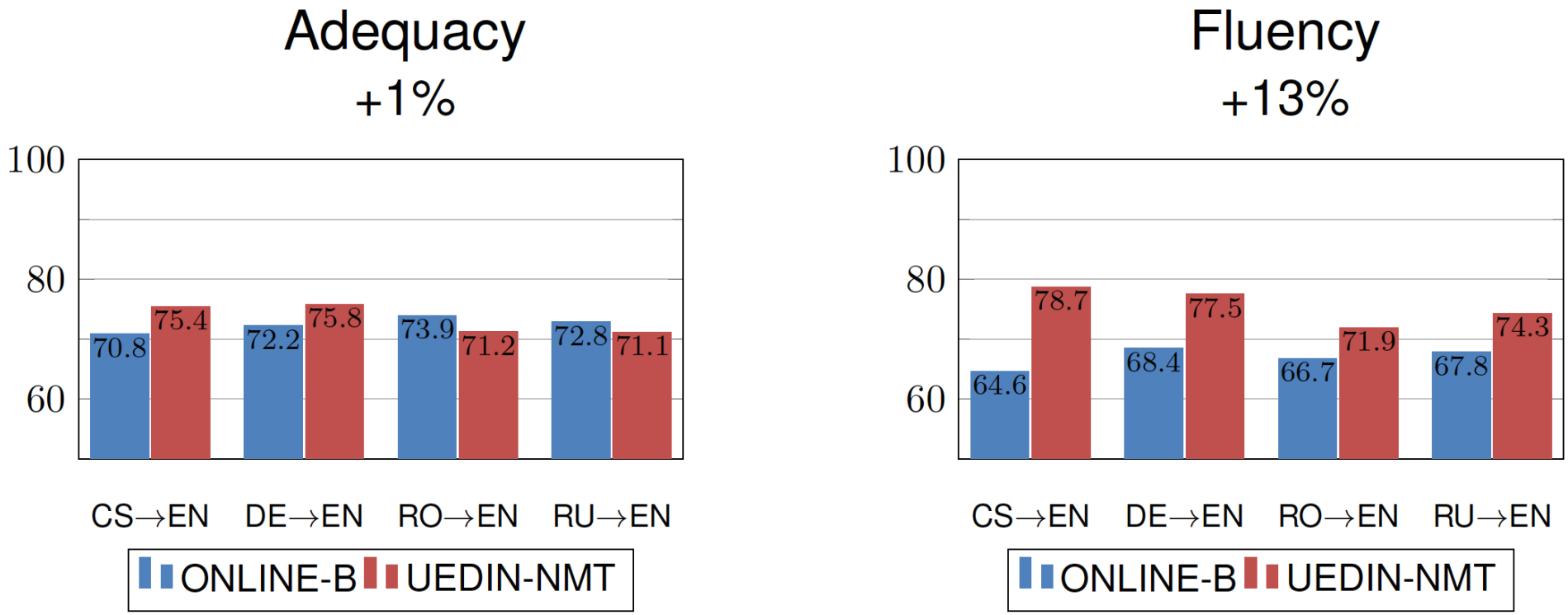
Researchers in Microsoft's labs in Beijing and in Redmond and Washington have developed an AI machine translation system that can translate with the same accuracy as a human from Chinese to English.

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

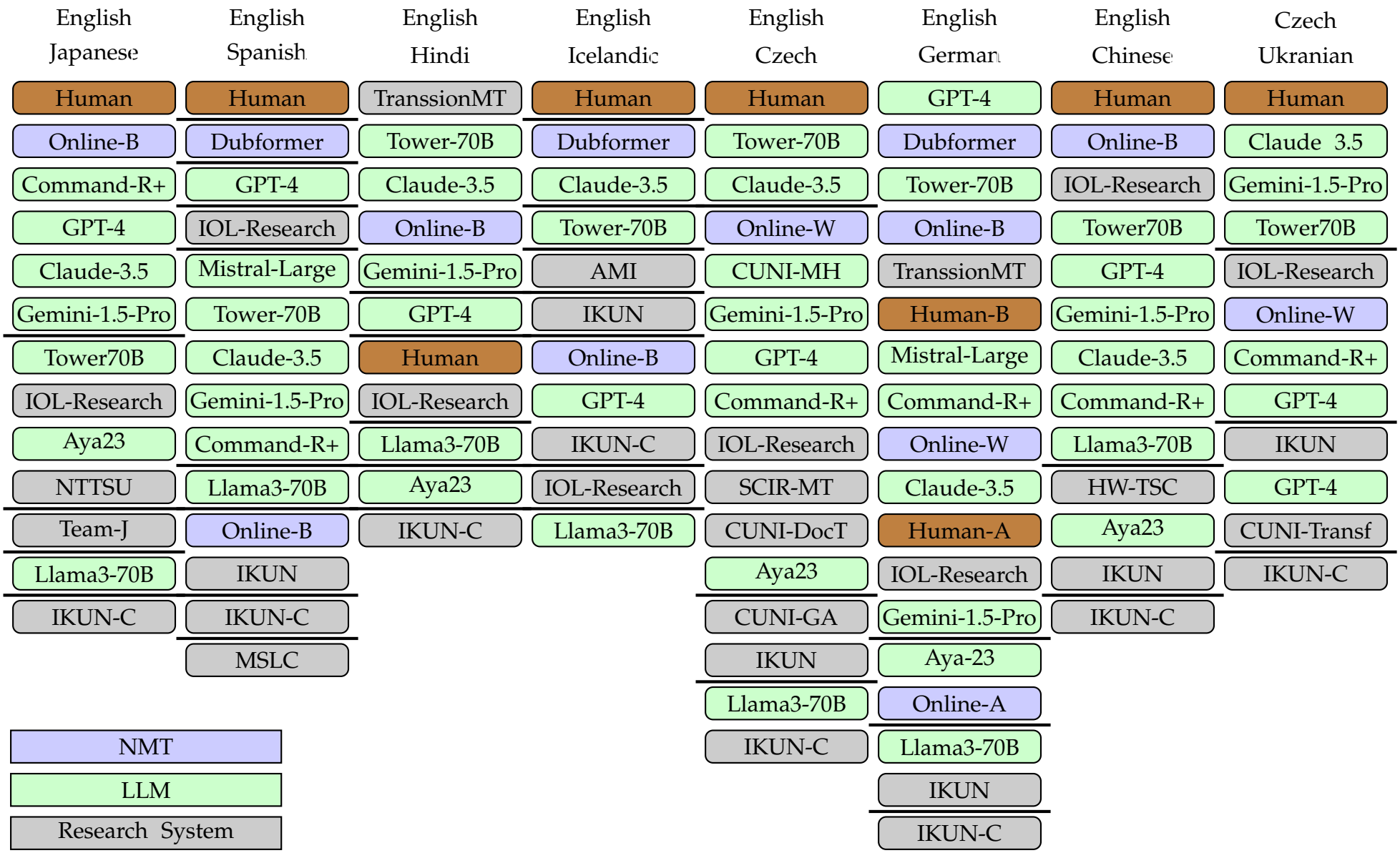
"90% of the system's output labelled as perfect by professional Russian-English translators"

Just Better Fluency?



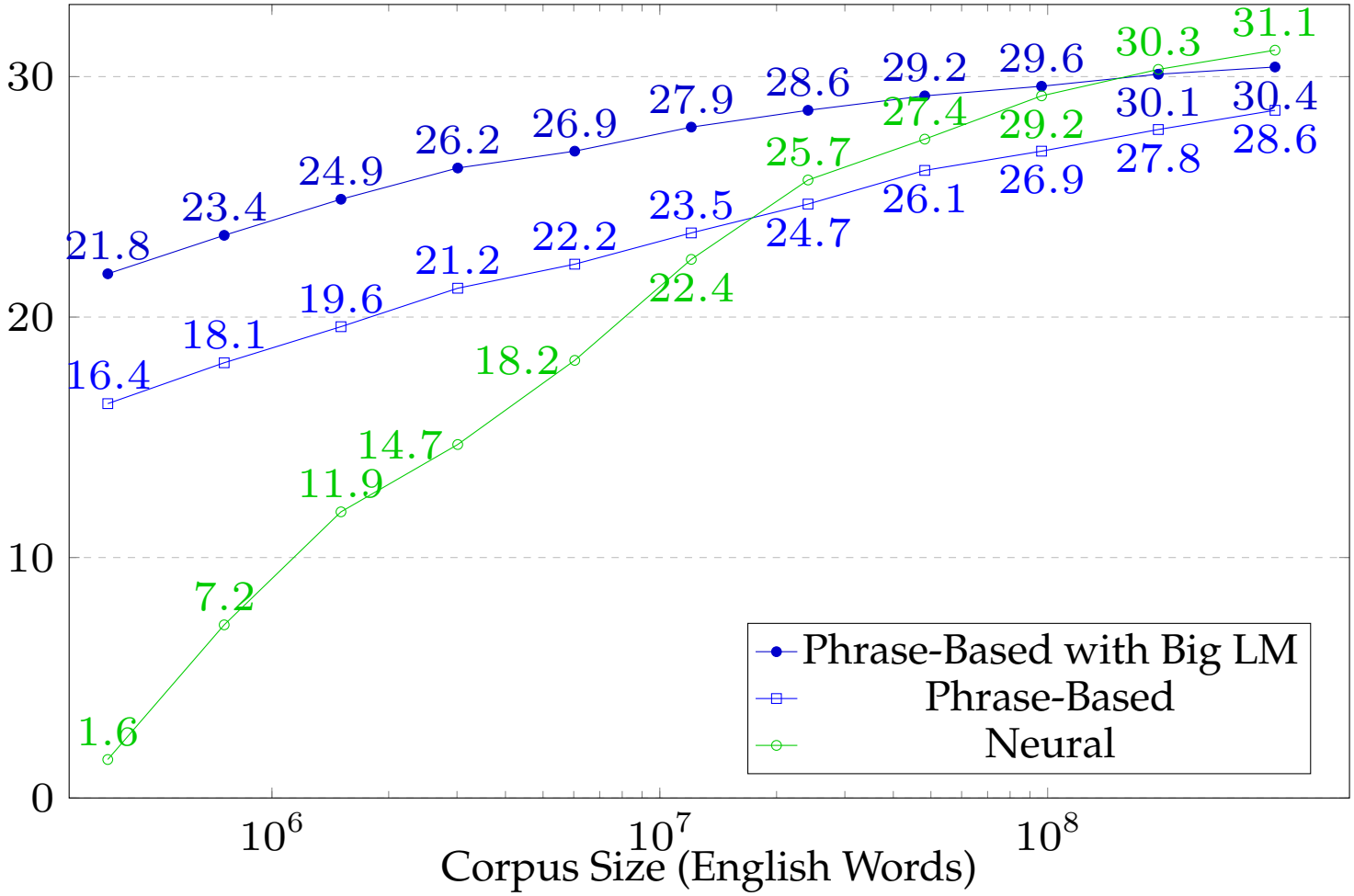
(from: Sennrich and Haddow, 2017)

WMT 2024: LLMs Arrive



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

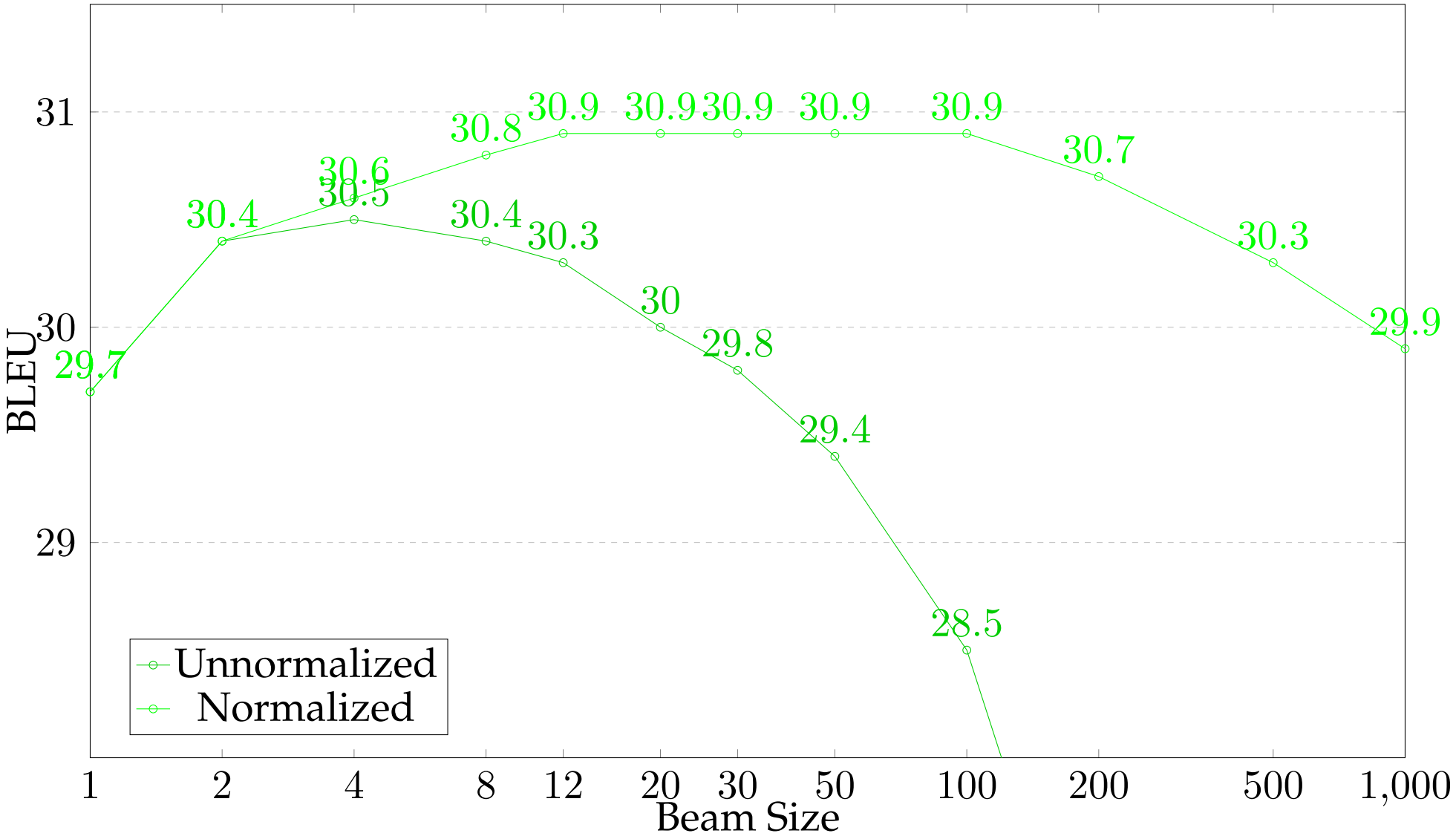
System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 32.8	45.1 42.2	35.3 44.7	17.9 17.9	26.4 20.8
Law	31.1 34.4	12.1 18.2	3.5 6.9	1.3 2.2	2.8 6.0
Medical	3.9 10.2	39.4 43.5	2.0 8.5	0.6 2.0	1.4 5.8
IT	1.9 3.7	6.5 5.3	42.1 39.8	1.8 1.6	3.9 4.7
Koran	0.4 1.8	0.0 2.1	0.0 2.3	15.9 18.8	1.0 5.5
Subtitles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


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

control over output

Specifying Decoding Constraints

- Overriding the decisions of the decoder
- Why?

  ⇒ translations have followed strict terminology

  ⇒ 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."

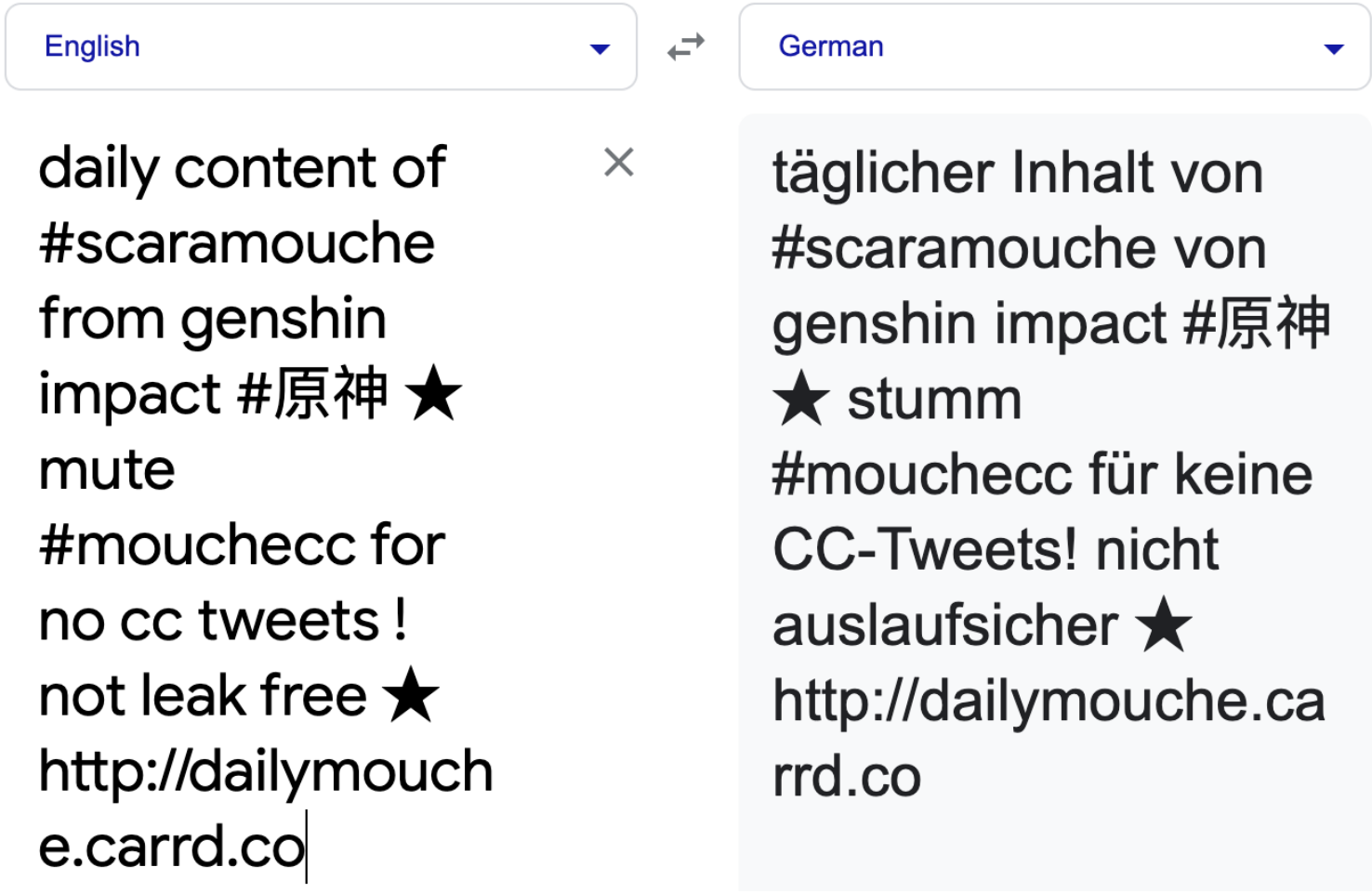
What are Catastrophic Errors?

- 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

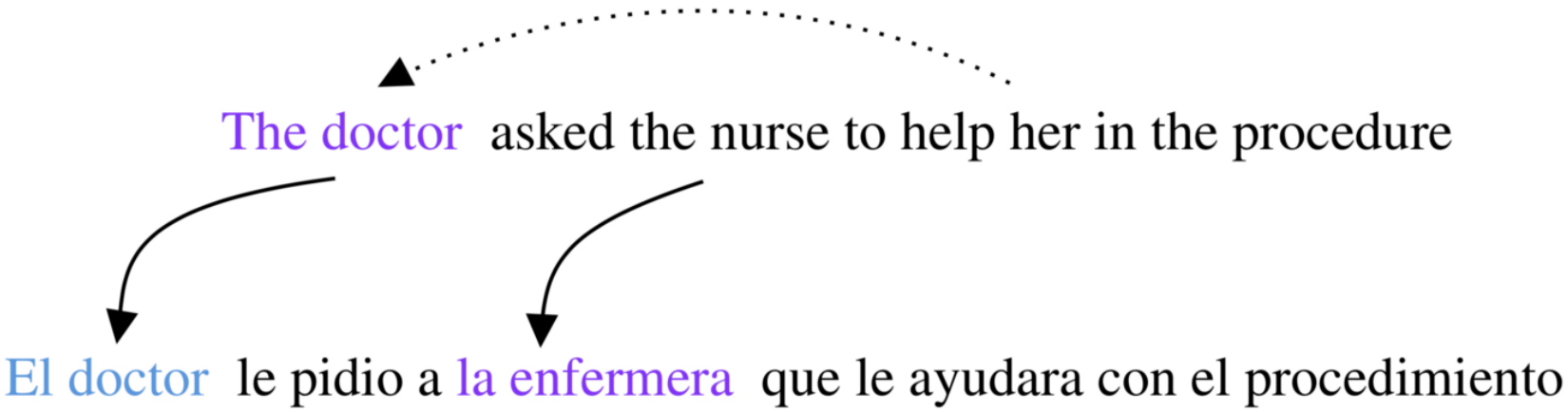
- Jargon and acronyms
- Misspellings (sometimes intended for effect)
- Mangled grammar
- Special symbols (emojis, etc.)
- Hashtags, URLs, ...
- Use of dialectical languages
- Use of non-standard writing systems (e.g., Latin script due to lack of keyboard)

Some Methods

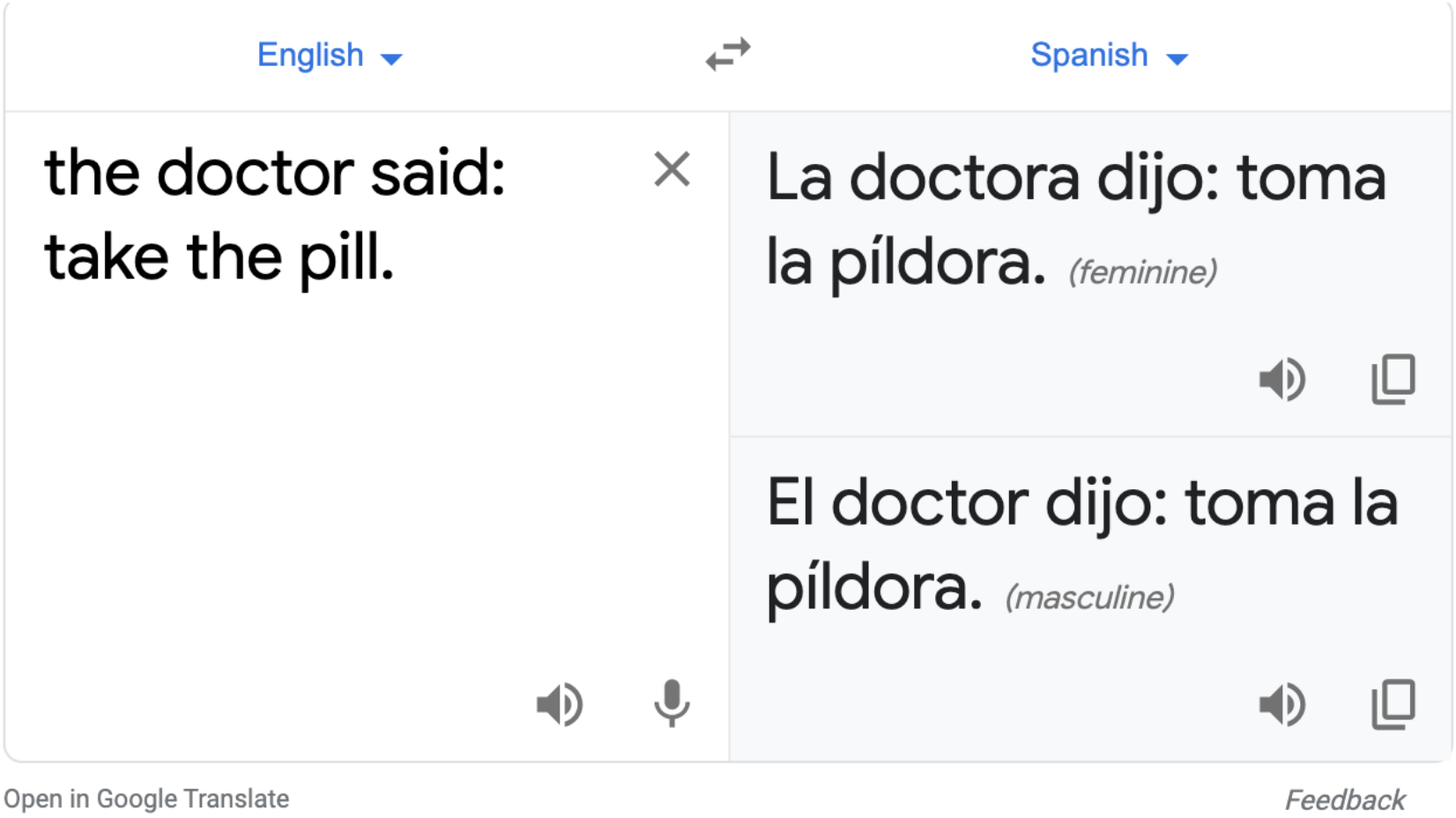
- Special handling of non-words like emojis, hashtags, URLs
- Creating synthetic noisy training data
- Adversarial training
- Resources
  - Machine translation of noisy text data set (MTNT)
  - WMT 2020 Shared Task on Machine Translation Robustness

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

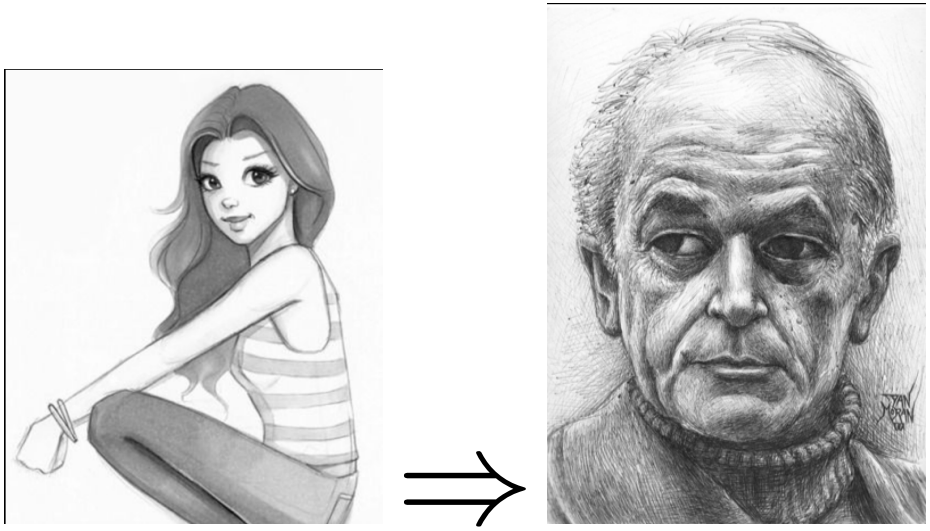
Gender Bias


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

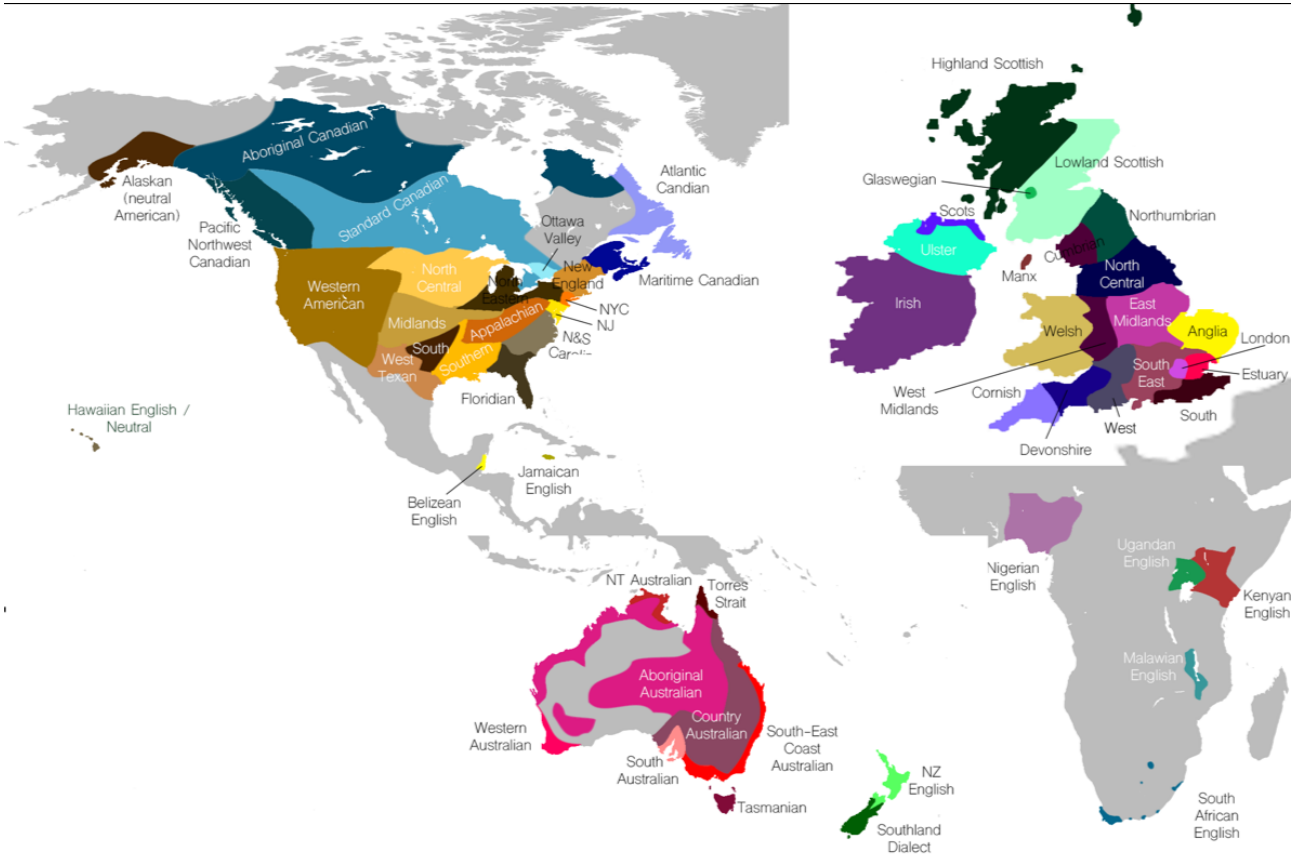
"You Sound Just Like Your Father"
Commercial Machine Translation Systems Include Stylistic Biases

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



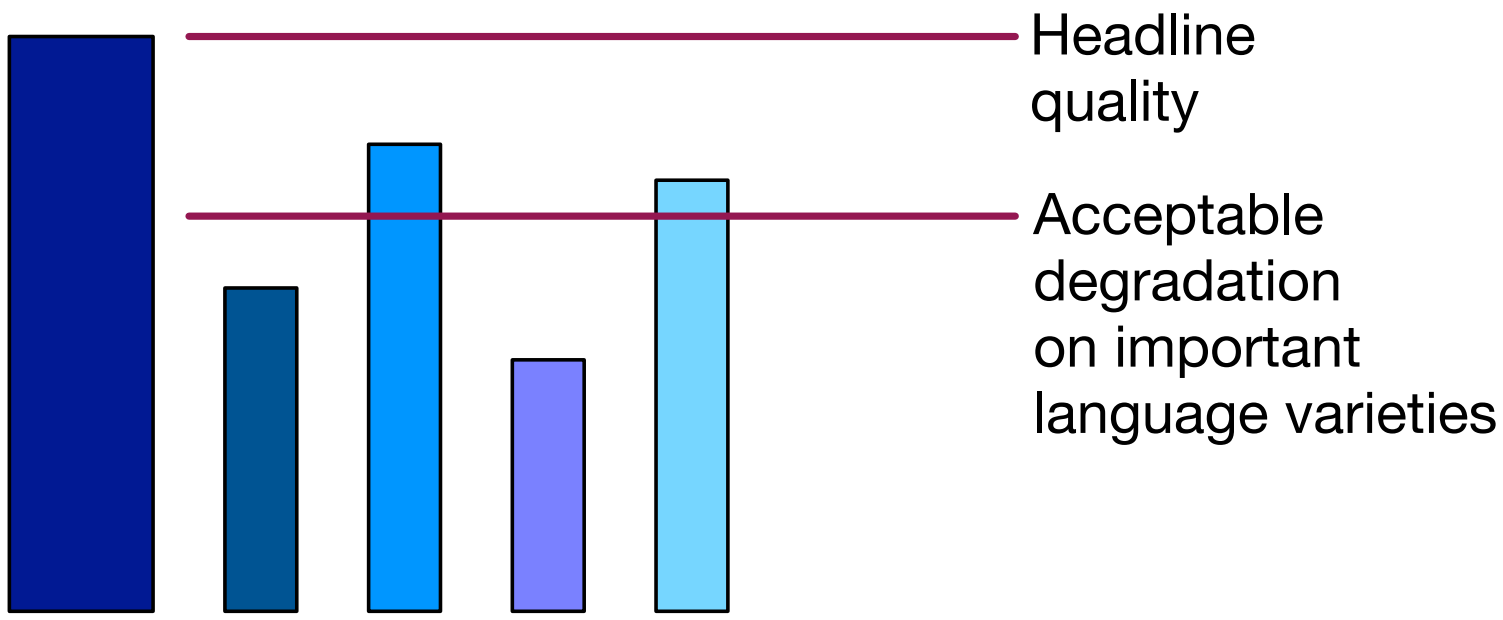
Dialect Bias

- 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.

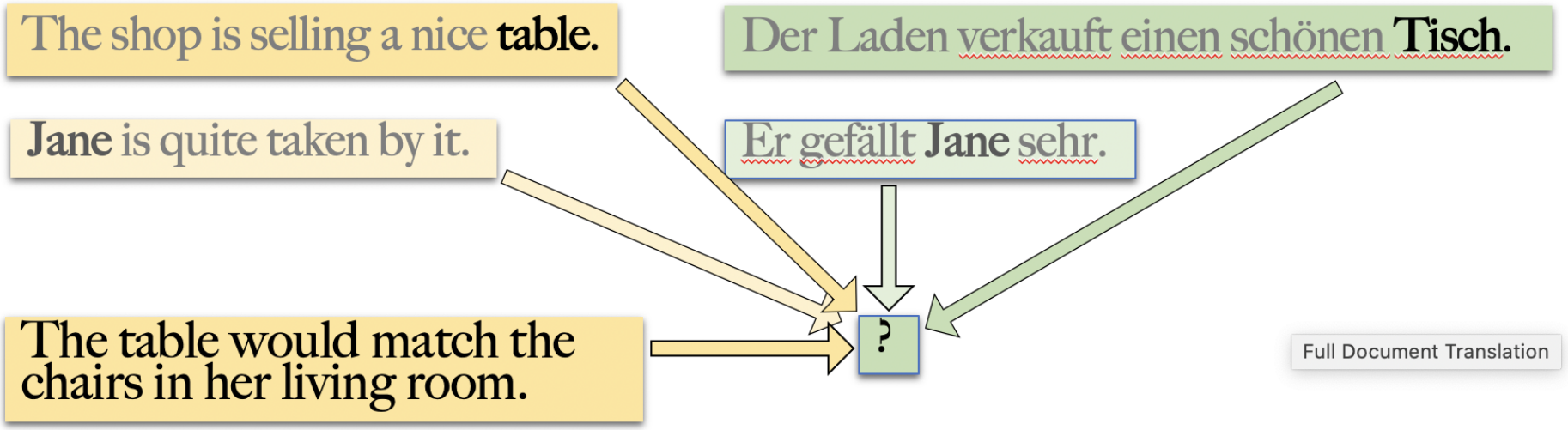
- Machine translation translates one sentence at a time
- But: surrounding context may help
  - translation of pronouns may require co-reference
  - ambiguous words may be informed by broader context

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.

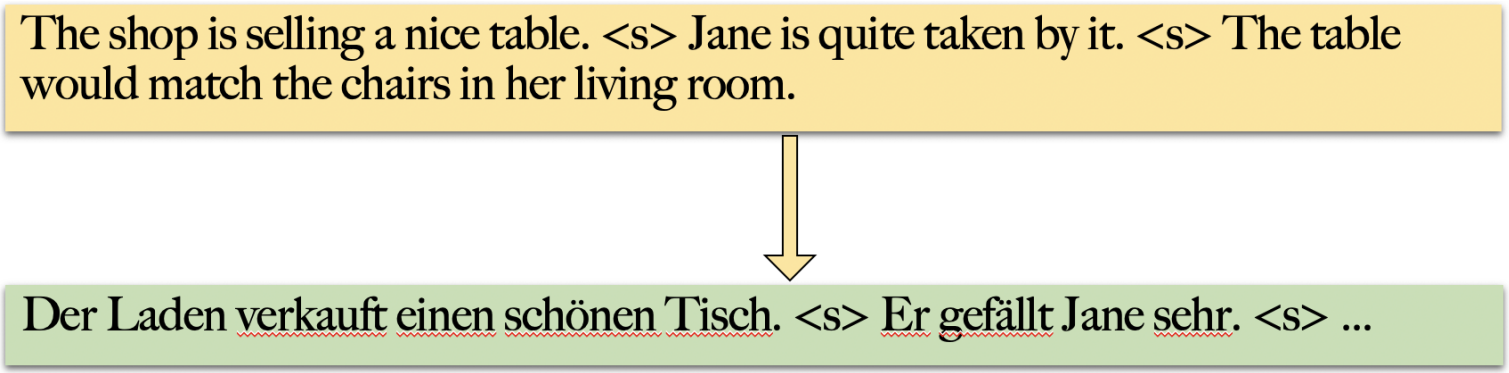
- Machine translation translates one sentence at a time
- But: surrounding context may help
  - translation of pronouns may require co-reference
  - ambiguous words may be informed by broader context
  - consistent translation of repeated words

Conditioning on Broader Context



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

Conditioning on Broader Context



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

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 +0.2 +0.2	26.6 24.2 -0.9 +0.2	24.7 24.4 -2.5 +0.4	20.9 24.8 -6.3 +0.8	17.3 25.2 -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 / -0.0	24.0 / -0.0	23.9 / -0.1	26.1 / -1.1 ; 23.9 / -0.1	25.3 / -1.9 ; 23.4 / -0.6

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

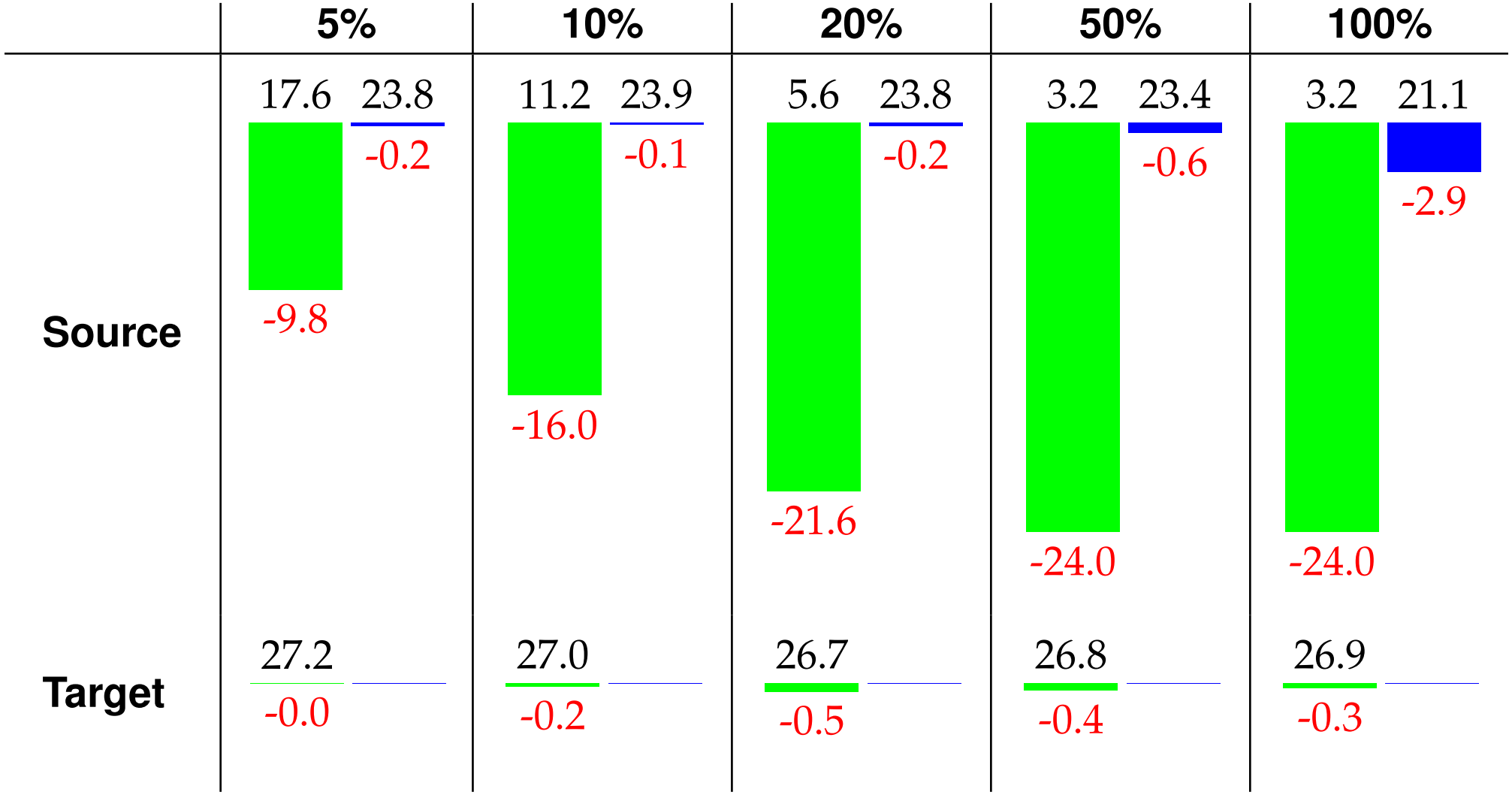
Misordered Words

- 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 24.0 -0.3 -0.0	26.8 23.9 -0.4 -0.1	26.8 23.9 -0.4 -0.1	26.8 23.9 -0.4 -0.1	26.8 23.8 -0.4 -0.2
fr target	26.7 24.0 -0.5 -0.0	26.6 23.9 -0.6 -0.1	26.7 23.8 -0.5 -0.2	26.2 23.5 -1.0 -0.5	25.0 23.4 -2.2 -0.6

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

Short Sentences

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

- No harm done

noise filtering

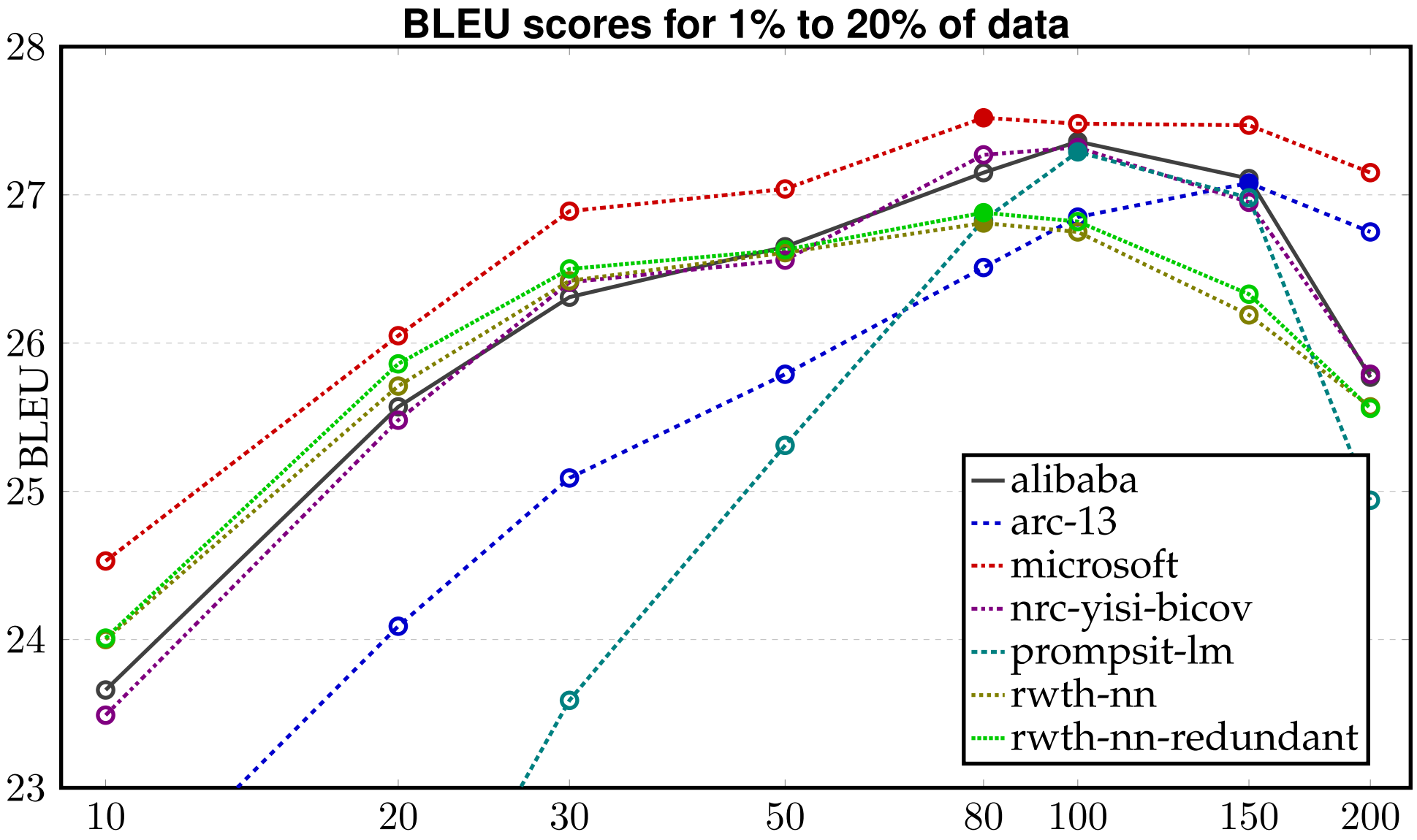
Detecting Noisy Training Data

- Noisy data is bad

$\Rightarrow$ remove it from the training data

- Shared Task Setup
  - given very noise web crawled parallel corpus (1 billion words)
  - to do: assign each sentence pair a quality score
  - training NMT model on fixed-sized subsets based on score
  - evaluation of translation quality of the system

Results (WMT 2018)



Dual Cross Entropy Filtering

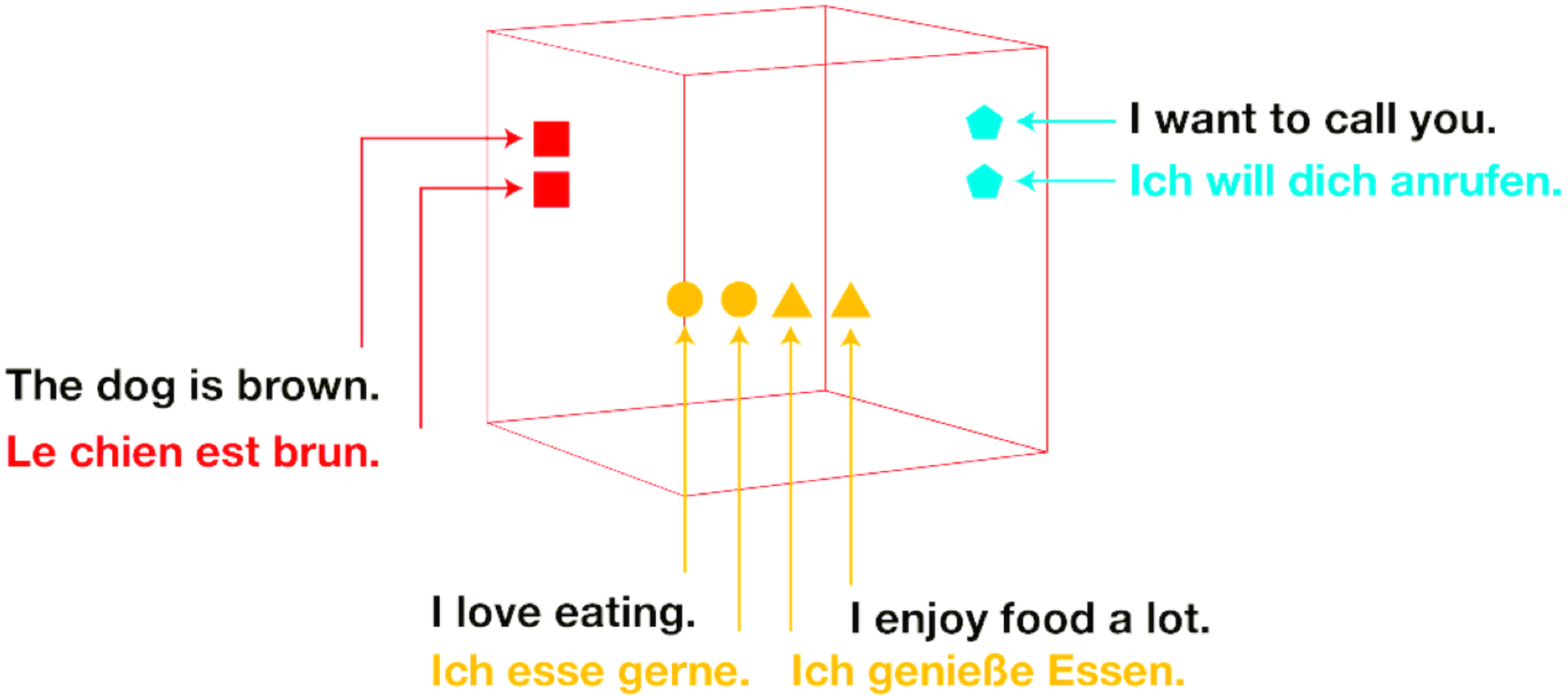
- Train translation model on clean data in both directions [Junczys-Dowmunt, 2018]
- Force-decode the target side of the sentence pair
  - compute path cost based on word prediction probabilities

$$H(y|x) = -\frac{1}{|y|}\sum_{t=1}^{|y|}\log P_M(y_t|y_{<t}, x)$$

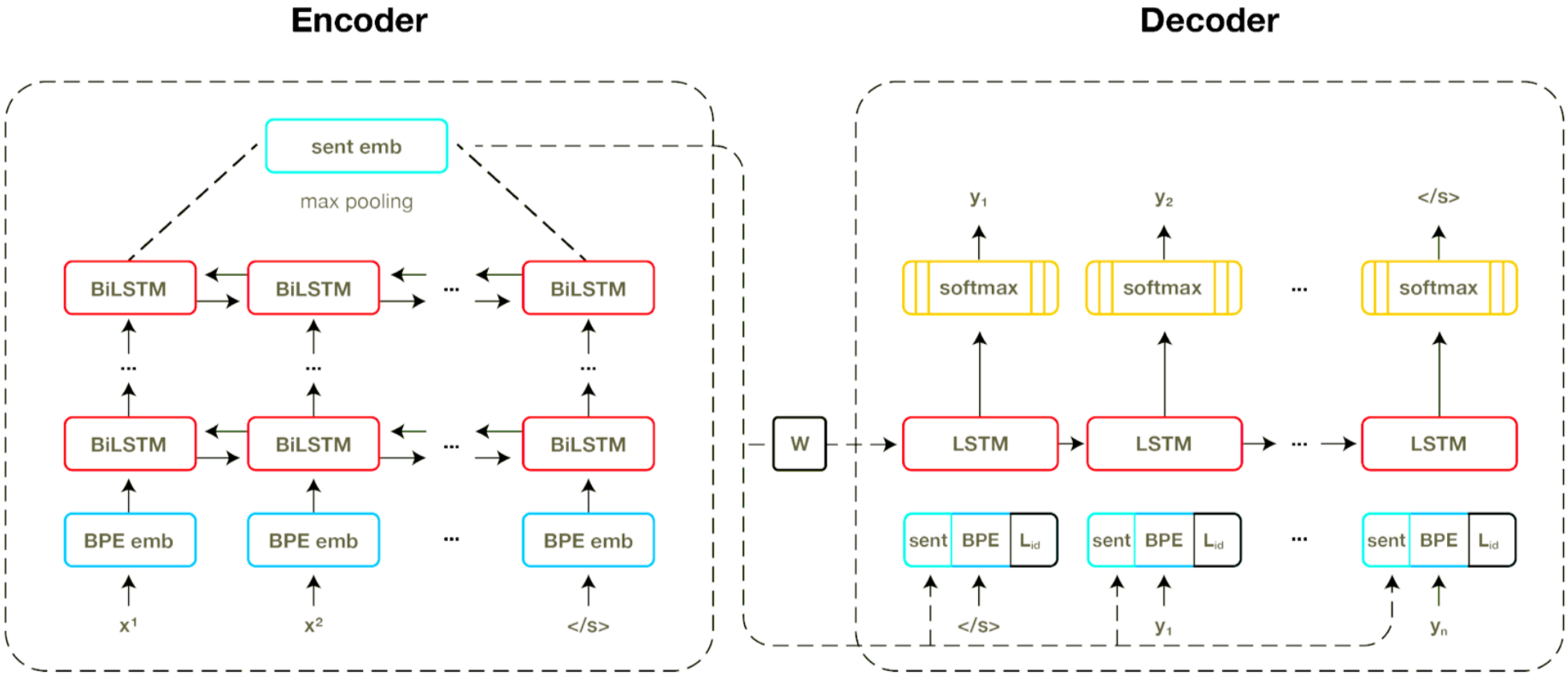
  - both directions: $H_A(x|y)$ and $H_B(y|x)$
- Combine scores

$$\underbrace{|H_A(x|y) - H_B(y|x)|}_{\text{similar costs}} + \underbrace{\frac{1}{2}(H_A(x|y) + H_B(y|x))}_{\text{low average cost}}$$

Cross-Lingual Sentence Embeddings

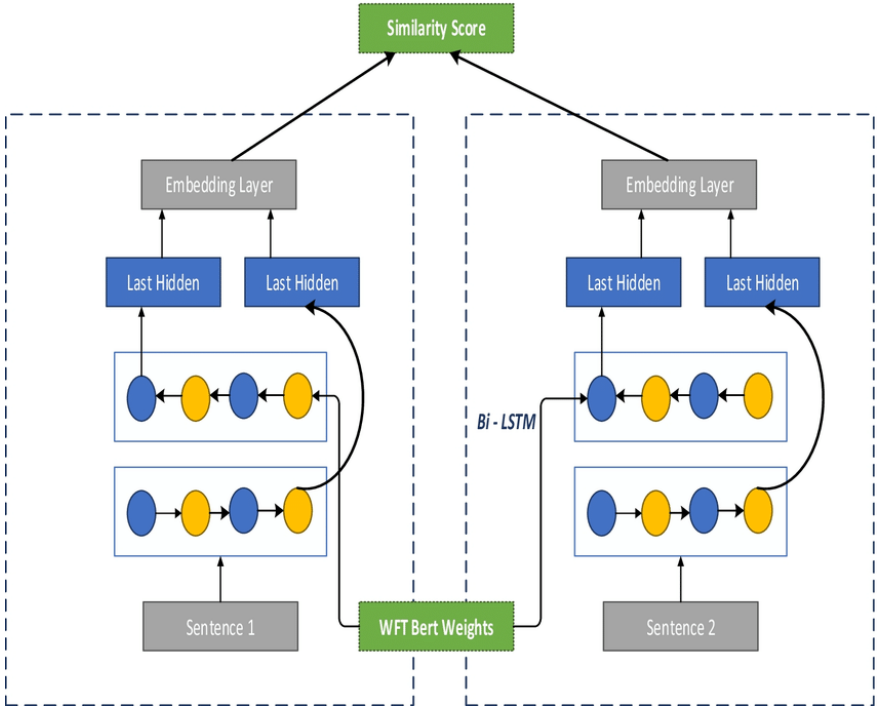


Cross-Lingual Sentence Embeddings



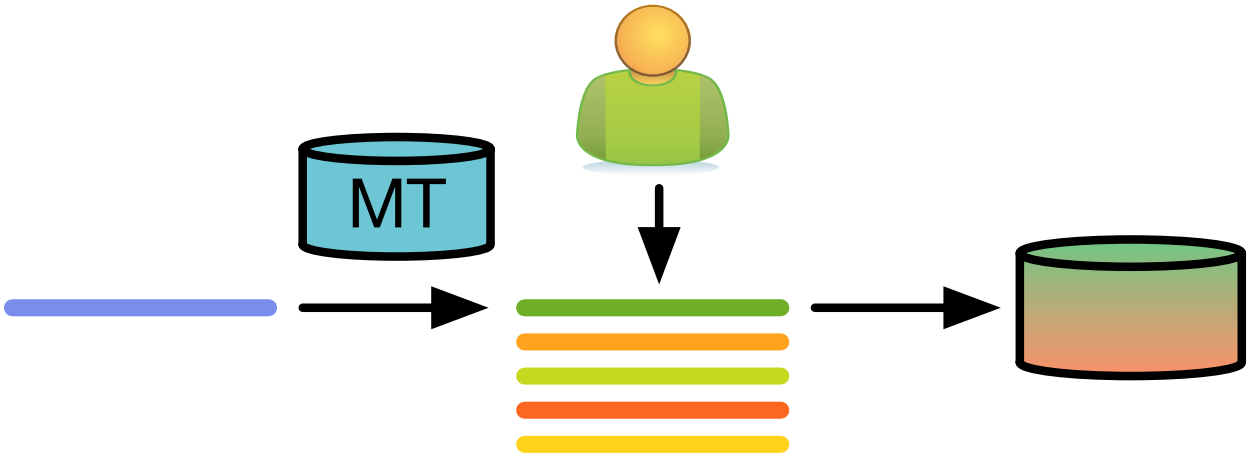
- LASER: Neural machine translation model with bottleneck feature

Training a Classifier



- Supervised learning problem (Siamese network)
  - good: sentence pair from known parallel corpus
  - bad: corrupted sentence pair (by changing some of the words or phrases)
- Important that the *bad* examples are not too bad
  → otherwise task too easy, not good model learned

Using Quality Estimation Models



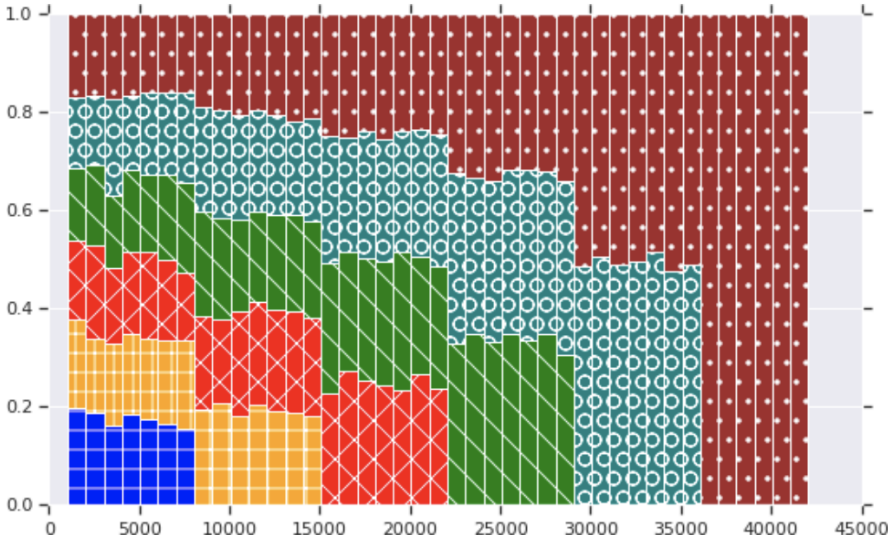
- Quality estimation models trained on user preference data (assessment of MT output)

- Assessing quality of a sentence pair essentially the same task

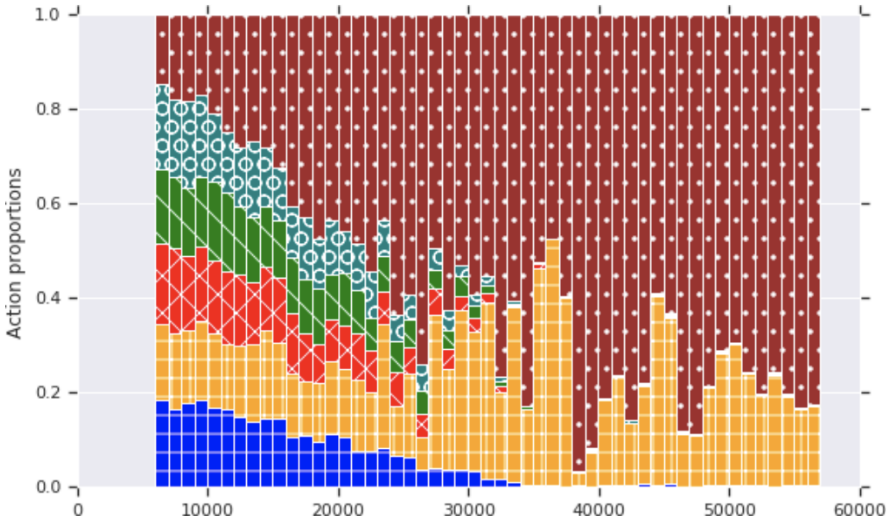
Curriculum Training

- Consider the order of training examples presented during training
  - easy to hard
  - general to in-domain
  - noisy to clean

- Order may be learning with reinforcement learning



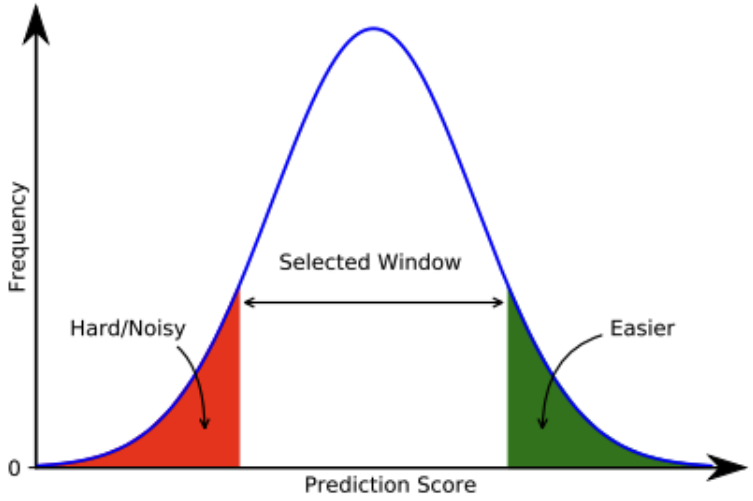
manual designed noisy-to-clean



reinforcement learning

[from my PhD student Kumar et al., 2019]

Ignoring Hard Examples During Training



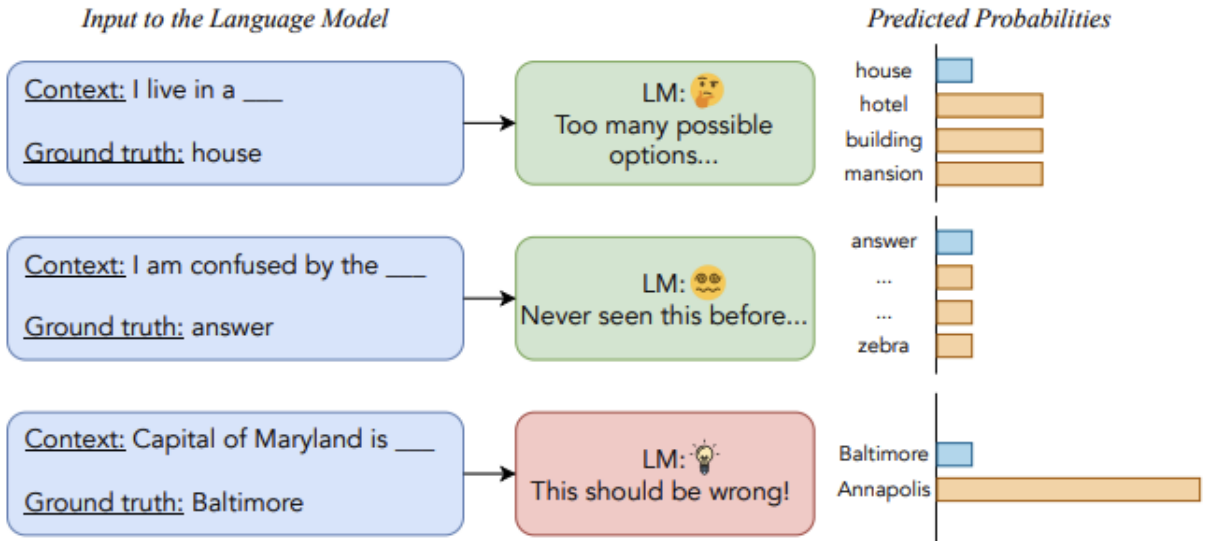
[from Mohiuddin et al., 2022]

- Motivation: Hard training samples will lead the model into bad directions
  - not far enough to handle the hard sample
  - corrupting model for more common samples
  - hard sample may be outlier (noise)

- Curriculum training: at even epoch (or sub-epoch) score all sentence pairs
  - ignore top 10% loss examples
  - ignore bottom 10% loss examples

Error Norm Truncation

- Considering this at the token level: large losses are a problem [Li et al., 2024]
- Solution: truncate large error norms: $||p_\theta(|y_{<t}, x) - \text{ONEHOT}(y_t)||_2$
- Examples:



- Last example is noise — we do not want to change the model (too much)
  → throw out its loss

Role of Data

- We are training text transformer models on diverse sets of data, e.g.,
  - relevant to language (pair) or not
  - relevant to domain or not
  - relevant to task or not
  - monolingual vs. parallel vs. task data
  - noise
  - synthesized vs. naturally occurring
- We have many techniques, e.g.,
  - over/undersampling, filtering
  - curriculum training: changes over time
  - behavior of samples during training process (e.g., current loss)
  - adaptation of subset of parameters or lower-dimensional parameter matrices
- Given the many choices, exhaustive experimentation is impossible (especially for training foundation models)

→ Currently a question of experience, best practices, rules of thumb

questions?