Online Versus Offline NMT Quality

An In-depth Analysis on English-German and German-English

Maha Elbayad[1,2] Michael Ustaszewski[3] Emmanuelle Esperança-Rodier[1]

Francis Brunet-Manquat[1] Jakob Verbeek[4] Laurent Besacier[1]

(1) Université Grenoble Alpes — LIG

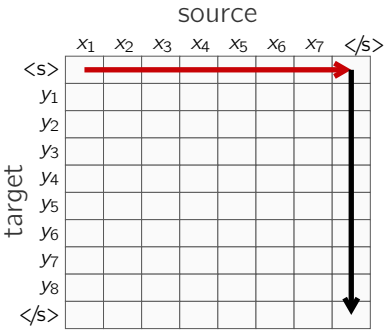
(2) Inria informatiques mathématiques

(3) universität innsbruck

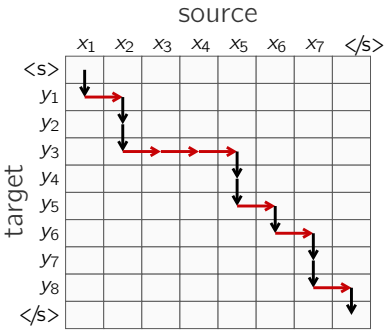
(4) FACEBOOK AI

Outline

Online Neural Machine Translation



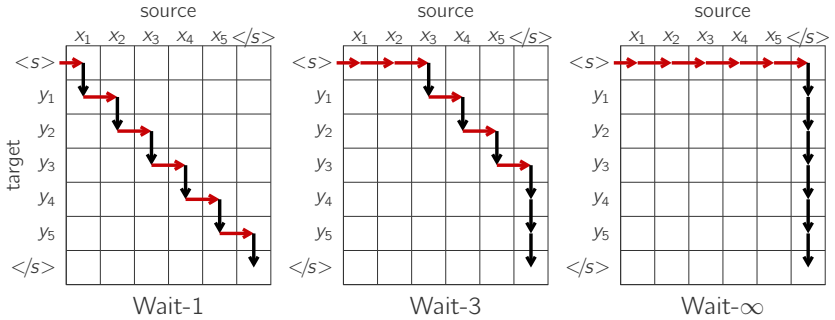
Offline translation



Online translation

Wait-k Decoders for Online Translation

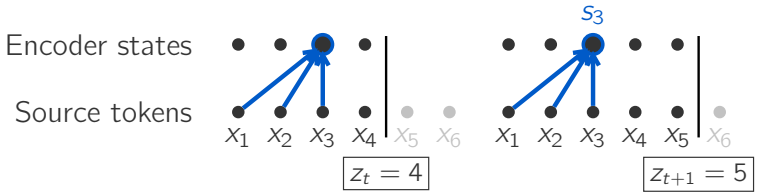
$$\forall t \in [1..|\mathbf{y}|], \quad z_t^{\text{wait-}k} = \min(k + t - 1, |\mathbf{x}|)$$



Wait-k or prefix-to-prefix decoding (Dalvi *et al.* 2018; Ma *et al.* 2019; Elbayad *et al.* 2020)

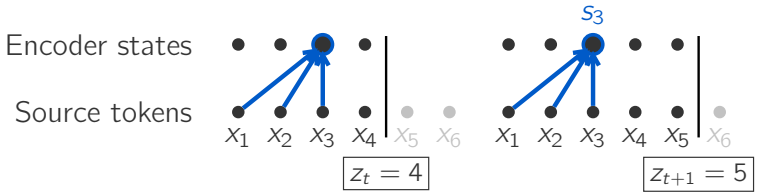
Online Transformer

- **Unidirectional encoder** (Elbayad *et al.* 2020)

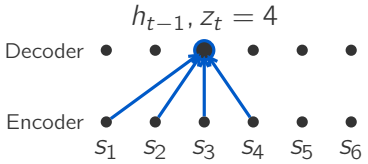


Online Transformer

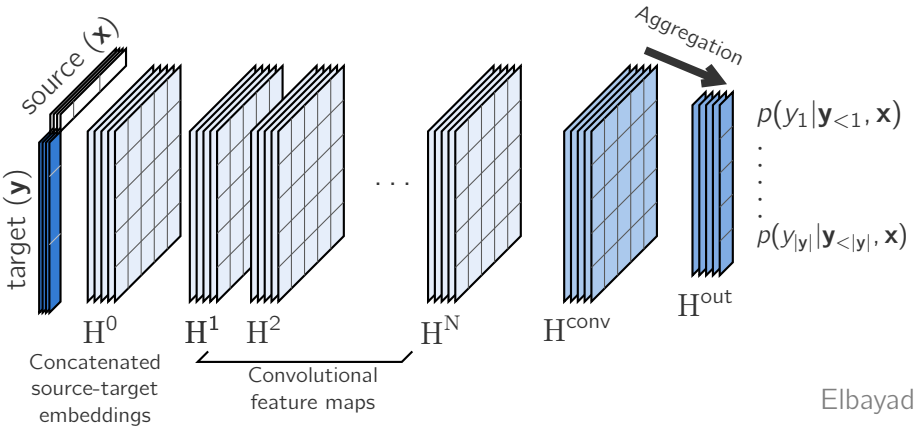
- **Unidirectional encoder** (Elbayad *et al.* 2020)



- **Masked decoder** - masking the attention energies wrt. $\mathbf{z}_t$

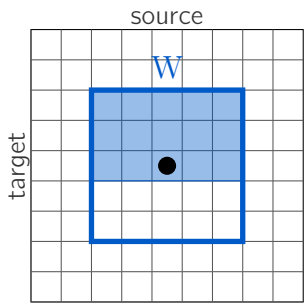


The Pervasive Attention Architecture

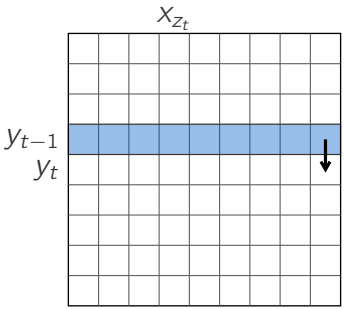


Elbayad *et al.* 2018

Online Pervasive Attention

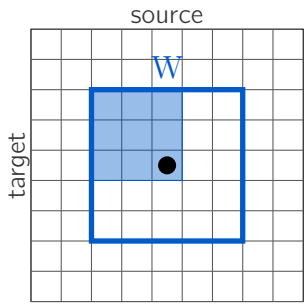


2D causal convolution

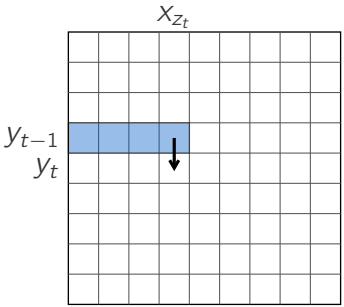


Features aggregation

Online Pervasive Attention



+ Masking the future source for unidirectional encoding.



The appropriate context size z_t is controlled during aggregation.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) → 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) → 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) → 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) → 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Models

- For each direction and for each architecture, an online and an offline model.
- Pervasive Attention (**PA**) with 14 layers and 7×7 filters (effectively 4×4).
- Transformer (**TF**) small.
- Online trained with $k_{\text{train}} = 7$ and evaluated with $k_{\text{eval}} = 3$.
- Greedy decoding for all.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Models

- For each direction and for each architecture, an online and an offline model.
- Pervasive Attention (**PA**) with 14 layers and 7×7 filters (effectively 4×4).
- Transformer (**TF**) small.
- Online trained with $k_{\text{train}} = 7$ and evaluated with $k_{\text{eval}} = 3$.
- Greedy decoding for all.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Models

- For each direction and for each architecture, an online and an offline model.
- Pervasive Attention (**PA**) with 14 layers and 7×7 filters (effectively 4×4).
- Transformer (**TF**) small.
- Online trained with $k_{\text{train}} = 7$ and evaluated with $k_{\text{eval}} = 3$.
- Greedy decoding for all.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Models

- For each direction and for each architecture, an online and an offline model.
- Pervasive Attention (**PA**) with 14 layers and 7×7 filters (effectively 4×4).
- Transformer (**TF**) small.
- Online trained with $k_{\text{train}} = 7$ and evaluated with $k_{\text{eval}} = 3$.
- Greedy decoding for all.

Training and Evaluation Setup

Data

- IWSLT'14 De-En and En-De (Cettolo *et al.* 2014).
- Sentences >175 words and pairs with length-ratio >1.5 are removed.
- The data is tokenized but not lowercased.
- The sequences are BPE segmented (Sennrich *et al.* 2016) $\rightarrow$ 32K vocabulary.
- Training = 160K, development = 7.3K and test = 6.7K.

Models

- For each direction and for each architecture, an online and an offline model.
- Pervasive Attention (**PA**) with 14 layers and 7×7 filters (effectively 4×4).
- Transformer (**TF**) small.
- Online trained with $k_{\text{train}} = 7$ and evaluated with $k_{\text{eval}} = 3$.
- Greedy decoding for all.

Metrics and Analysis Factors

- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.

Metrics and Analysis Factors

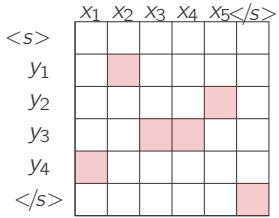
- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.
- Translation delay with AL (Ma *et al.* 2019)

Metrics and Analysis Factors

- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.
- Translation delay with AL (Ma *et al.* 2019)
- Source length $|\mathbf{x}|$.
  + Target-side and source-side relative positions in the paper.

Metrics and Analysis Factors

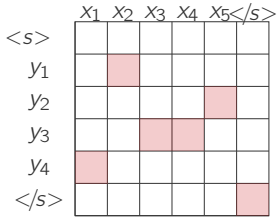
- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.
- Translation delay with AL (Ma *et al.* 2019)
- Source length $|\mathbf{x}|$.
  + Target-side and source-side relative positions in the paper.
- Lagging difficulty $\mathrm{LD}(\mathbf{x}, \mathbf{y})$.



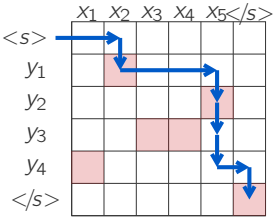
Alignments from fast-align

Metrics and Analysis Factors

- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.
- Translation delay with AL (Ma *et al.* 2019)
- Source length $|\mathbf{x}|$.
  + Target-side and source-side relative positions in the paper.
- Lagging difficulty $\mathrm{LD}(\mathbf{x}, \mathbf{y})$.



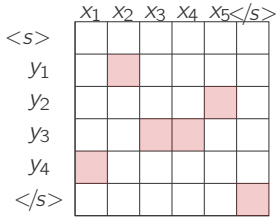
Alignments from fast-align



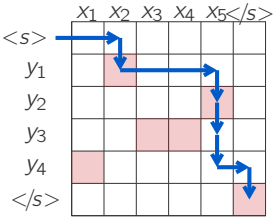
Estimated ideal path

Metrics and Analysis Factors

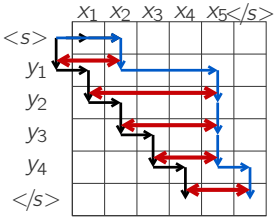
- Translation quality.
  BLEU (Papineni *et al.* 2002), METEOR (Lavie *et al.* 2007),
  TER (Snover *et al.* 2006). + ROUGE-L (Lin 2004) and BERTScore (Zhang *et al.* 2020) in the paper.
- Translation delay with AL (Ma *et al.* 2019)
- Source length $|\mathbf{x}|$.
  + Target-side and source-side relative positions in the paper.
- Lagging difficulty $\mathrm{LD}(\mathbf{x}, \mathbf{y})$.



Alignments from fast-align



Estimated ideal path



Average distance from wait-0

Automatic Evaluation

Bold = the better scoring.
Underlined = better than its competitor with at least 95% statistical significance.

	De→En		En→De	
	PA	TF	PA	TF
	Offline	Offline	Offline	Offline
↑BLEU	**31.24**	31.13	26.03	**26.60**
↑METEOR	28.95	**29.25**	38.81	**39.37**
↓TER	0.56	**0.56**	0.63	**0.62**
AL	21.10	21.10	20.71	20.71

Automatic Evaluation

Bold = the better scoring.
Underlined = better than its competitor with at least 95% statistical significance.

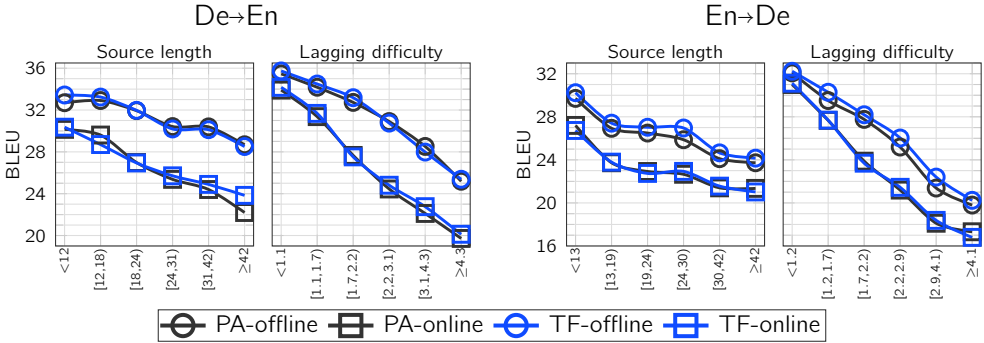
	De→En				En→De	
	PA	PA	TF	TF	PA	TF
	Offline	Online	Offline	Online	Offline	Offline
↑BLEU	**31.24**	26.44	31.13	**26.57**	26.03	**26.60**
↑METEOR	28.95	**25.97**	**29.25**	25.65	38.81	**39.37**
↓TER	0.56	**0.62**	**0.56**	0.64	0.63	**0.62**
AL	21.10	2.59	21.10	3.16	20.71	20.71

Automatic Evaluation

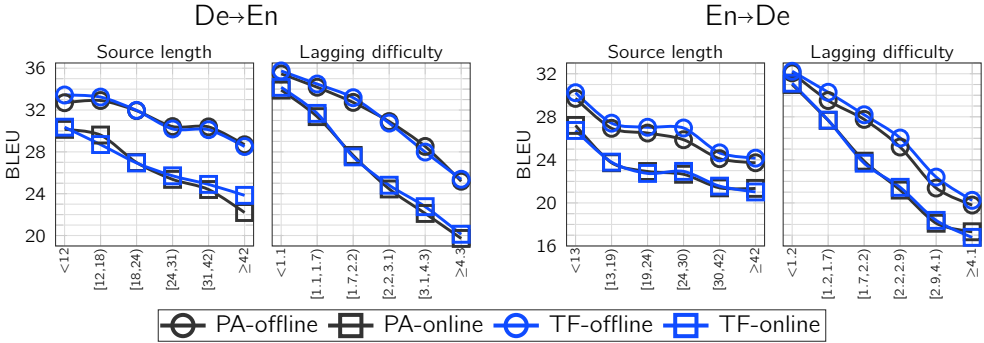
Bold = the better scoring.
Underlined = better than its competitor with at least 95% statistical significance.

	De→En				En→De			
	PA	PA	TF	TF	PA	PA	TF	TF
	Offline	Online	Offline	Online	Offline	Online	Offline	Online
↑BLEU	**31.24**	26.44	31.13	**26.57**	26.03	**23.04**	**26.60**	22.98
↑METEOR	28.95	**25.97**	**29.25**	25.65	38.81	**35.72**	**39.37**	35.35
↓TER	0.56	**0.62**	**0.56**	0.64	0.63	**0.68**	**0.62**	0.69
AL	21.10	2.59	21.10	3.16	20.71	3.33	20.71	3.49

Automatic Evaluation



Automatic Evaluation



Lagging difficulty (LD) is highly correlated with the translation quality (BLEU) in both online and offline translation tasks.

Human Evaluation

Annotation Setup

- 200 segments per language pair.

  $Q_1 \leq |\mathbf{x}| \leq Q_3$, equally sampled over lagging difficulty bins.

- ACCOLÉ (Esperança-Rodier *et al.* 2019).

  Web interface for annotation of error spans and types in source and target.

- Annotators.

  2 native translation experts per pair, annotation training on calibration set.

- Inter-annotator agreement compatible with other MQM-based studies.

  Cohen's κ = 0.33 (De→En) and 0.40 (En→De).

Annotation Setup

- 200 segments per language pair.

  $Q_1 \leq |\mathbf{x}| \leq Q_3$, equally sampled over lagging difficulty bins.

- ACCOLÉ (Esperança-Rodier *et al.* 2019).

  Web interface for annotation of error spans and types in source and target.

- Annotators.

  2 native translation experts per pair, annotation training on calibration set.

- Inter-annotator agreement compatible with other MQM-based studies.

  Cohen's κ = 0.33 (De→En) and 0.40 (En→De).

Annotation Setup

- 200 segments per language pair.

  $Q_1 \leq |\mathbf{x}| \leq Q_3$, equally sampled over lagging difficulty bins.

- ACCOLÉ (Esperança-Rodier *et al.* 2019).

  Web interface for annotation of error spans and types in source and target.

- Annotators.

  2 native translation experts per pair, annotation training on calibration set.

- Inter-annotator agreement compatible with other MQM-based studies.

  Cohen's $\kappa = 0.33$ (De→En) and 0.40 (En→De).

Annotation Setup

- 200 segments per language pair.

  $Q_1 \leq |\mathbf{x}| \leq Q_3$, equally sampled over lagging difficulty bins.

- ACCOLÉ (Esperança-Rodier *et al.* 2019).

  Web interface for annotation of error spans and types in source and target.

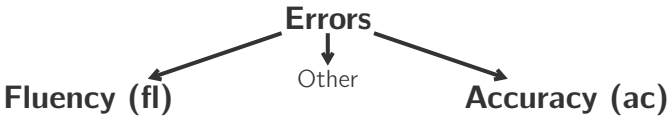
- Annotators.

  2 native translation experts per pair, annotation training on calibration set.

- Inter-annotator agreement compatible with other MQM-based studies.

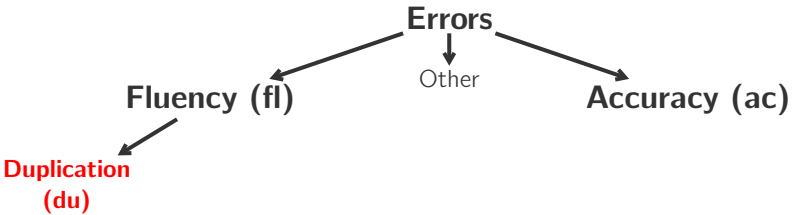
  Cohen's κ = 0.33 (De→En) and 0.40 (En→De).

Error Typology



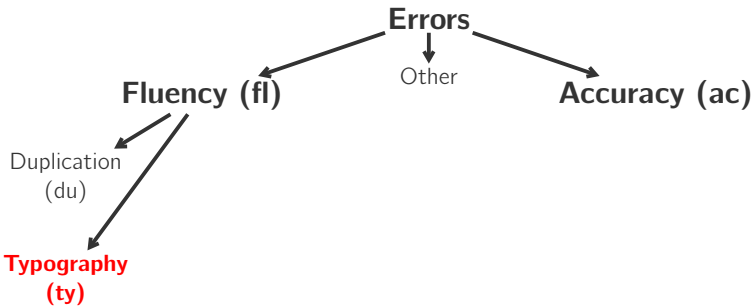
A pilot annotation was carried out to select a subset of the MQM error typology (Lommel *et al.* 2014) relevant to this study

Error Typology



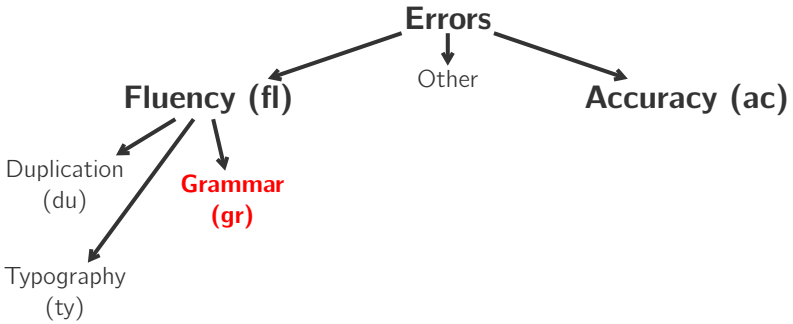
Reference: In high school, a classmate once said [...]
Hypothesis: In high school, once in high school, a fellow told us [...]

Error Typology



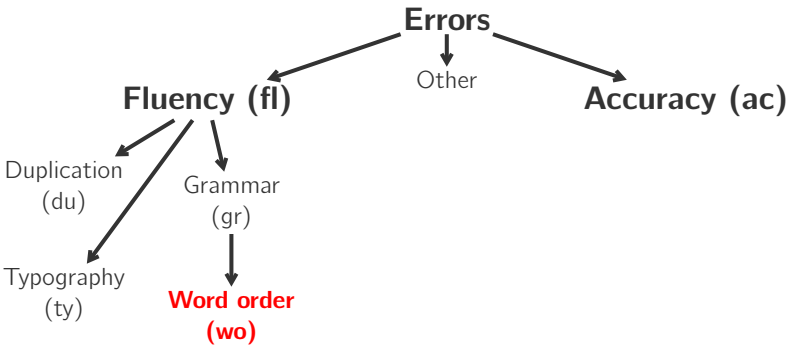
Reference: [...] meaning of the word "educate" comes from the root word "educe."
Hypothesis: [...] meaning of the word "educate" is rooted in the word "educe.

Error Typology



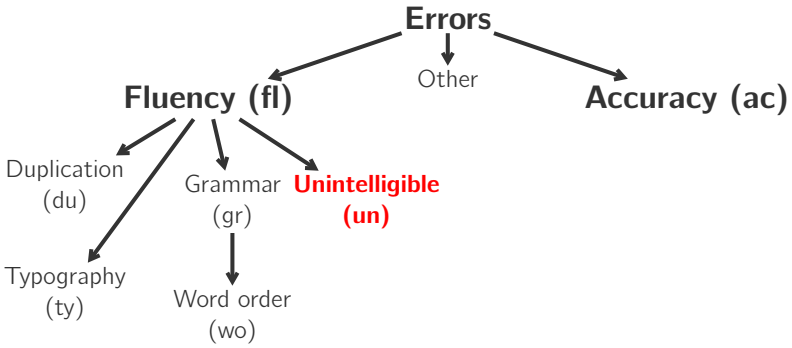
Reference: [. . .] you want to design things as intuitively as possible.
Hypothesis: [. . .] you want to make things so much intuitively possible.

Error Typology



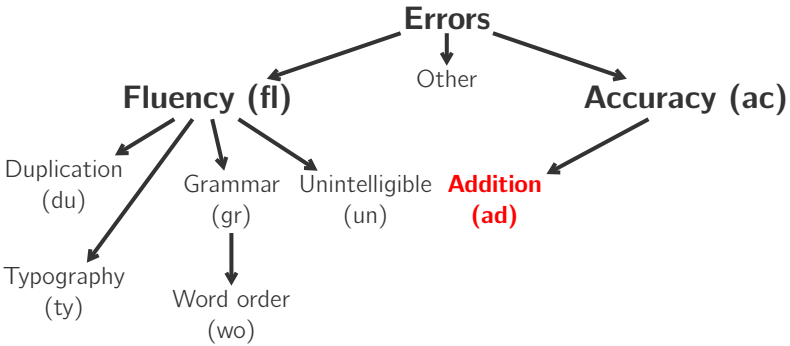
Reference: I was given another gift, which was to be able to see into the future [...]
Hypothesis: I got another gift, which is in the future to see [...].

Error Typology



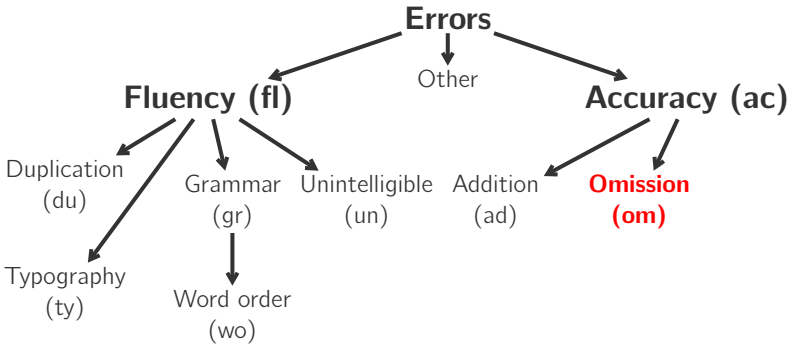
Reference: And for them language had inferior importance.
Hypothesis: What the language undergeals from subordination in the other time was.

Error Typology



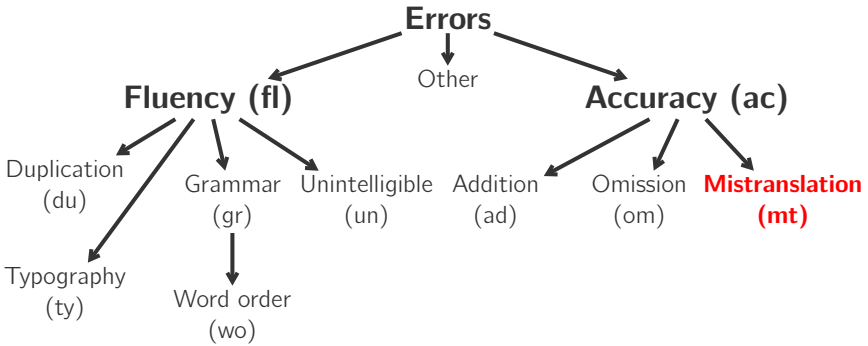
Reference: A couple months went by, and I had just forgotten all about it.
Hypothesis: A few months ago, I was going to go over, and I just forgot everything.

Error Typology



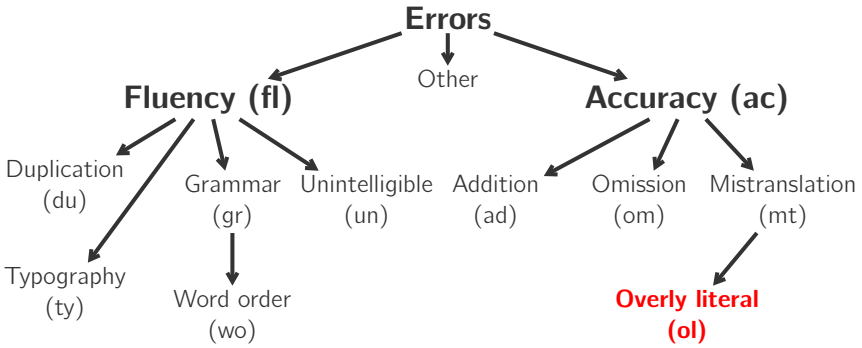
Reference: And if we can do this for raw data, why not do it for content as well?
Hypothesis: And if we do this for raw data, why not content itself?

Error Typology



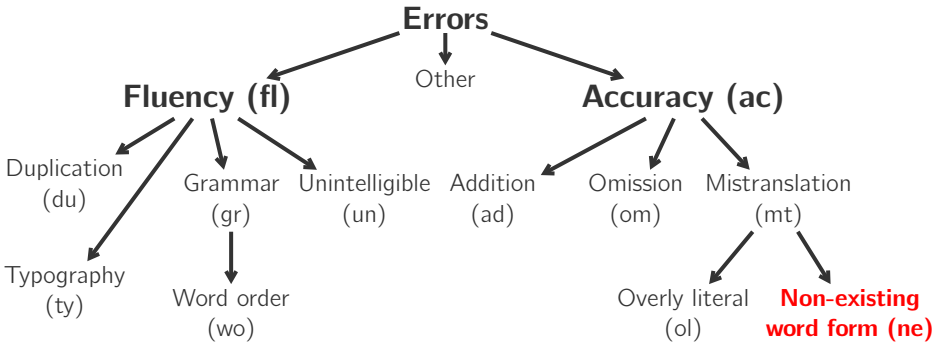
Reference: [. . .] I immediately went to look up the 2009 online edition [. . .]
Hypothesis: [. . .] I immediately started to call up the online copy of 2009 [. . .]

Error Typology



Reference: Three months later I had relocated [. . .]
Hypothesis: Three months later, I was moved [. . .]

Error Typology



Reference: the Tanzanian giraffe.
Hypothesis: the tansanic giraffe

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt) Mistranslation	**433**	587	+36	457	**572**	+25	245	260	+6	**202**	260	+29
(ne) Non-existing WF	26	17	-35	**14**	**16**	+14	**39**	58	+49	43	**54**	+26
(om) Omission	**67**	**113**	+69	96	127	+32	**99**	**74**	-25	126	114	-10
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt) Mistranslation	**433**	587	+36	457	**572**	+25	245	260	+6	**202**	260	+29
(ne) Non-existing WF	26	17	-35	**14**	**16**	+14	**39**	58	+49	43	**54**	+26
(om) Omission	**67**	**113**	+69	96	127	+32	**99**	**74**	-25	126	114	-10
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25

- Overall, **mistranslation** is the largest contributor to accuracy errors.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt) Mistranslation	**433**	587	+36	457	**572**	+25	245	260	+6	**202**	260	+29
(ne) Non-existing WF	26	17	-35	**14**	**16**	+14	**39**	58	+49	43	**54**	+26
(om) Omission	**67**	**113**	+69	96	127	+32	**99**	**74**	-25	126	114	-10
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25

- **Addition** errors increase drastically in En→De online.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt) Mistranslation	**433**	587	+36	457	**572**	+25	245	260	+6	**202**	260	+29
(ne) Non-existing WF	26	17	-35	**14**	**16**	+14	**39**	58	+49	43	**54**	+26
(om) Omission	**67**	**113**	+69	96	127	+32	**99**	**74**	-25	126	114	-10
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25

- De→En is more prone to **omissions** errors in the online setup.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl) Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du) Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr) Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty) Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(wo) Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+) Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76

- NMT systems are more prone to **accuracy** errors than **fluency** errors.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

		De→En						En→De					
		PA			TF			PA			TF		
	System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad)	Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt)													
(ne)													
(om)													
(ol)	Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+)	Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl)	Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du)	Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr)	Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty)	Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(wo)	Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+)	Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76

- More grammar errors are found in En→De compared to De→En.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl) Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du) Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr) Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty) Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(wo) Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+) Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76

- **Addition**, **word order** and **duplication** increase the most in the online setup.

Human Evaluation

Bold = The system (PA or TF) with fewer errors either online or offline.

	De→En						En→De					
	PA			TF			PA			TF		
System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad) Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt) Mistranslatio											260	+29
(ne) Non-existing											**54**	+26
(om) Omission											114	-10
(ol) Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+) Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl) Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du) Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr) Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty) Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(wo) Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+) Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76

- **Duplication** is extremely problematic for TF.

Human Evaluation

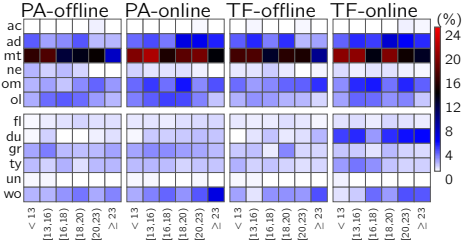
Bold = The system (PA or TF) with fewer errors either online or offline.

		De→En						En→De					
		PA			TF			PA			TF		
	System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ad)	Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt)	Mistr												+29
(ne)	Non-												+26
(om)	Omis												-10
(ol)	Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+)	Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl)	Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du)	Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr)	Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty)	Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(wo)	Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+)	Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76

- PA is more prone to **grammar** and **overly literal** errors.

Human Evaluation

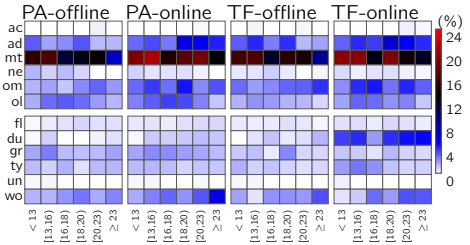
Source length De→En



Mistranslation errors peak in short segments due to the ease of spotting accuracy errors in fluent short segments.

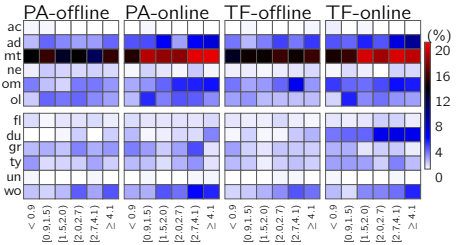
Human Evaluation

Source length De→En



Mistranslation errors peak in short segments due to the ease of spotting accuracy errors in fluent short segments.

Lagging difficulty De→En



Addition and **omission** errors are particularly correlated with LD.

Conclusion

- We ran the first human evaluation of offline and online NMT systems for spoken language translation.
- We highlighted the weaknesses of wait-k Transformer (duplication) and Pervasive Attention (grammar, overly literal).
- We introduced a strong indicator in lagging difficulty that is highly correlated with translation quality, particularly in online translation.
- Our annotated data is made available at `https://github.com/elbayadm/OnlineMT-Evaluation`

References I

- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L. & Federico, M. *Report on the 11th IWSLT evaluation campaign.* in *Proc. of IWSLT* (2014).
- Dalvi, F., Durrani, N., Sajjad, H. & Vogel, S. *Incremental Decoding and Training Methods for Simultaneous Translation in Neural Machine Translation.* in *Proc. of NAACL-HLT* (2018).
- Elbayad, M., Besacier, L. & Verbeek, J. *Efficient Wait-k Models for Simultaneous Machine Translation.* in *Proc. of INTERSPEECH* (2020).
- Elbayad, M., Besacier, L. & Verbeek, J. *Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction.* in *Proc. of CoNLL* (2018).

References II

- Esperança-Rodier, E., Brunet-Manquat, F. & Eady, S. *ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora.* in *Translating and the computer 41* (2019).
- Lavie, A. & Agarwal, A. *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments.* in *Proc. of the Second Workshop on Statistical Machine Translation* (June 2007).
- Lin, C.-Y. *ROUGE: A Package for Automatic Evaluation of Summaries.* in *Text Summarization Branches Out* (2004).
- Lommel, A., Uszkoreit, H. & Burchardt, A. A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica* (2014).

References III

- Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H. & Wang, H. *STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework.* in *Proc. of ACL* (2019).
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. *BLEU: a Method for Automatic Evaluation of Machine Translation.* in *Proc. of ACL* (2002).
- Sennrich, R., Haddow, B. & Birch, A. *Neural Machine Translation of Rare Words with Subword Units.* in *Proc. of ACL* (2016).
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. *A study of translation edit rate with targeted human annotation.* in *Proc. of AMTA* (2006).

References IV

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. *Attention Is All You Need.* in *Proc. of NeurIPS* (2017).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. *Bertscore: Evaluating text generation with bert.* in *Proc. of ICLR* (2020).