Adapting Neural Machine Translation with Parallel Synthetic Data

Mara Chinea-Ríos, Álvaro Peris, Francisco Casacuberta
{machirio, lvapeab, fcn}@prhlt.upv.es

Pattern Recognition and Human Language Technologies
Research Center
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València

DSIC

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Outline

Introduction: reminder of automatic MT evaluation

- Most common automatic MT metrics:
  - BiLingual Evaluation Understudy (BLEU)
  - Translation Edit Rate (TER)
  - METEOR
- We usually have the output from a MT system h and a reference(s) to compare: $\hat{y}$.

TER

$$TER = \frac{\#edits}{|\hat{y}|}$$

- $edits$: Edit operations applied to h for transforming it into $\hat{y}$.
  - Additions.
  - Deletions.
  - Substitutions.
  - Shifts of word sequences.

- $\#edits$ is obtained by dynamic programming.

BLEU

- Modified n-gram precision
  - n-gram precision:

$$C(n_t) = \frac{n_t}{|h|}$$

  - n_t: n-grams $\in h$ that appear in $\hat{y}$.

- Clipped n-gram precision:

$$C_{clip}(n_t) = \min(C(n_t), n_{t_{max}})$$

$n_{t_{max}}$: Maximum total count of n_t in any of the reference translations.

BLEU

- Steps for computing BLEU:
  1. Compute n-gram matches sentence by sentence ($\mathcal{C}$).
  2. Add the clipped n-gram counts for all these candidate sentences.
  3. Divide this result by the number of candidate n-grams in the test corpus.

$$p_n = \frac{\sum_{\mathcal{C} \in \{Hypotheses\}} \sum_{n_t \in \mathcal{C}} C_{clip}(n_t)}{\sum_{\mathcal{C}' \in \{Hypotheses\}} \sum_{n'_t \in \mathcal{C}'} C(n'_t)}$$

BLEU

- Brevity penalty (BP) factor: Penalizes short translations:

$$BP = \begin{cases} 1, & \text{if } |h| > |\hat{y}| \\ e^{1-\frac{|\hat{y}|}{|h|}}, & \text{if } |h| \leq |\hat{y}| \end{cases}$$

BLEU

- BLEU: weighted geometric mean of the n-grams for combining them.

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- Typically:
  - $N = 4$
  - $w_n = \frac{1}{N}$

Problems with BLEU

- Recall is not measured in BLEU.
  - Recall: proportion of matched n-grams out of the total number of n-grams in $\hat{y}$.
- The order of the n-grams may be small.
- Lack of explicit word matching between $\hat{y}$ and h.
  - This can result in incorrect matchings (e.g. function words).

METEOR

1. Align unigrams from $\hat{y}$ and h. A valid alignment must map each unigram with at most one unigram in the other string

   (a) Compute all possible mappings:
   - Exact mapping: `computer` $\neq$ `computers`.
   - Stemmed mapping: `computer` $=$ `computers`.
   - Synonym mapping: `computer` $=$ `PC`

   (b) Select the largest valid subset of unigrams from (a).

2. Compute unigram precision (P) and unigram recall R.
   - P: $\frac{\#\text{unigrams in h that are mapped in } \hat{y}}{|h|}$
   - R: $\frac{\#\text{unigrams in h that are mapped in } \hat{y}}{|y|}$

METEOR

3. Compute harmonic $FMean$ of P and R:

$$Fmean = \frac{10PR}{R + 9P}$$

4. Compute an alignment penalty: Group all unigrams in h that are mapped in $\hat{y}$ into the fewest number of chunks.

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

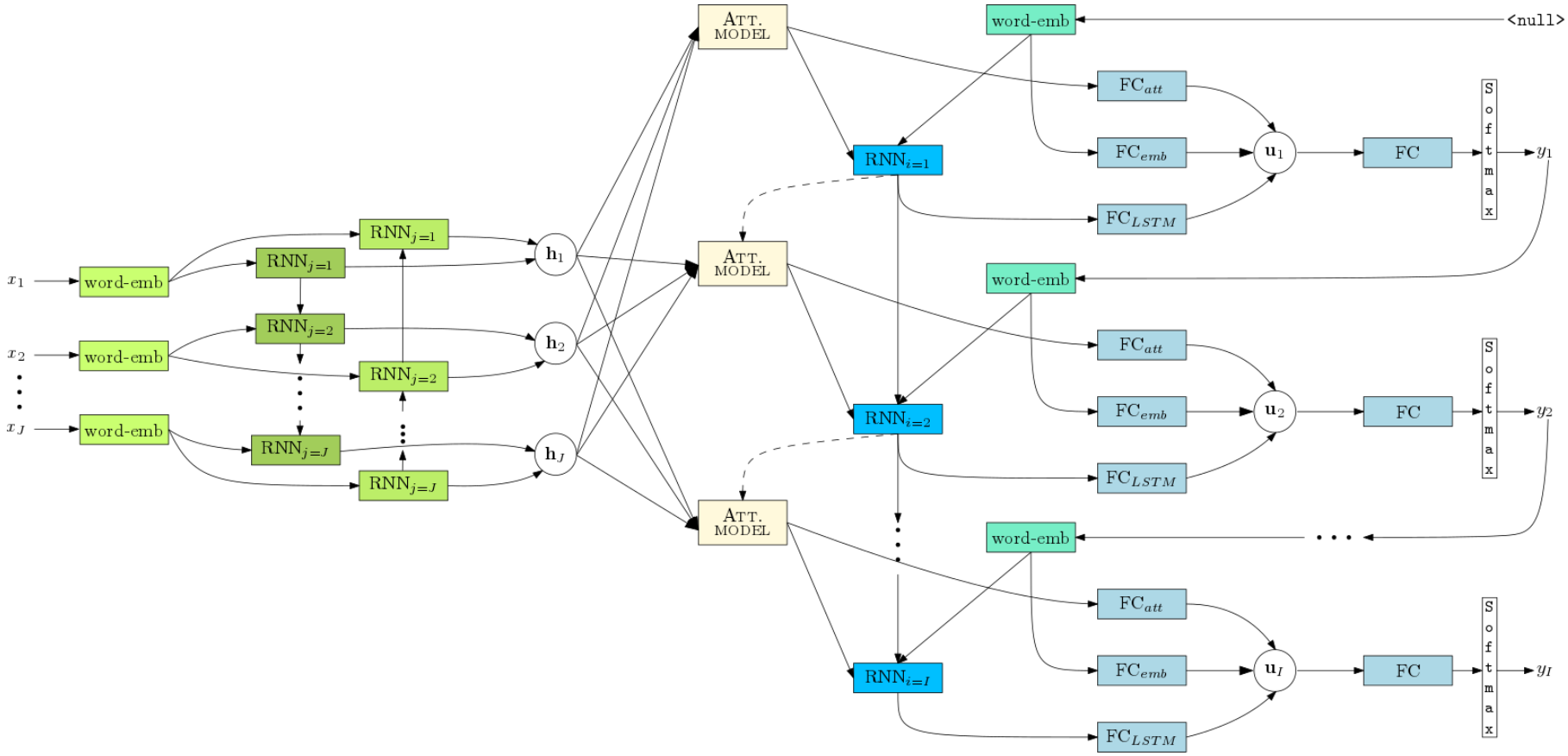
METEOR

5. Compute Meteor $Score$ for a given alignment:

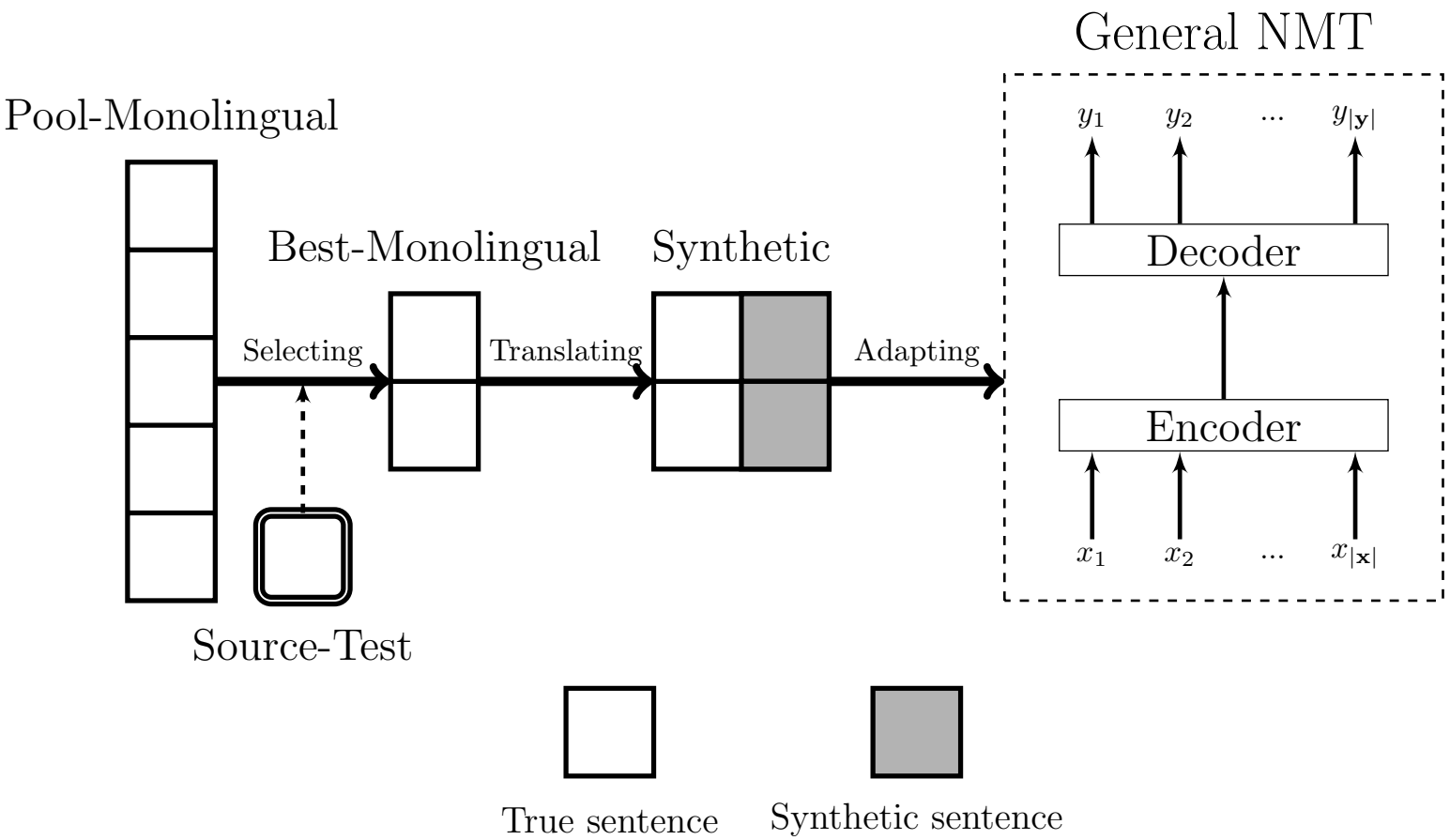
$$Score = FMean * (1 - Penalty)$$

6. Overall METEOR score for a given system: Aggregate all scores and combine them using the same formulas.

Neural machine translation



Synthetic creation method



Training procedure

- NMT systems:
  - Byte Pair Encoding ($32k$ merge operations).
  - LSTM networks.
  - LSTM, word embedding and attention sizes: 512.
  - Maximum likelihood training.
  - Adam (learning rate: 0.0002).
  - $L2$ norm of gradients clipped to 1.
  - Layer normalization and Gaussian noise ($\sigma = 0.01$).

- Standard configuration of Moses:
  - 5-gram language model (KN-discount).
  - Symmetrised word alignments.
  - MERT.

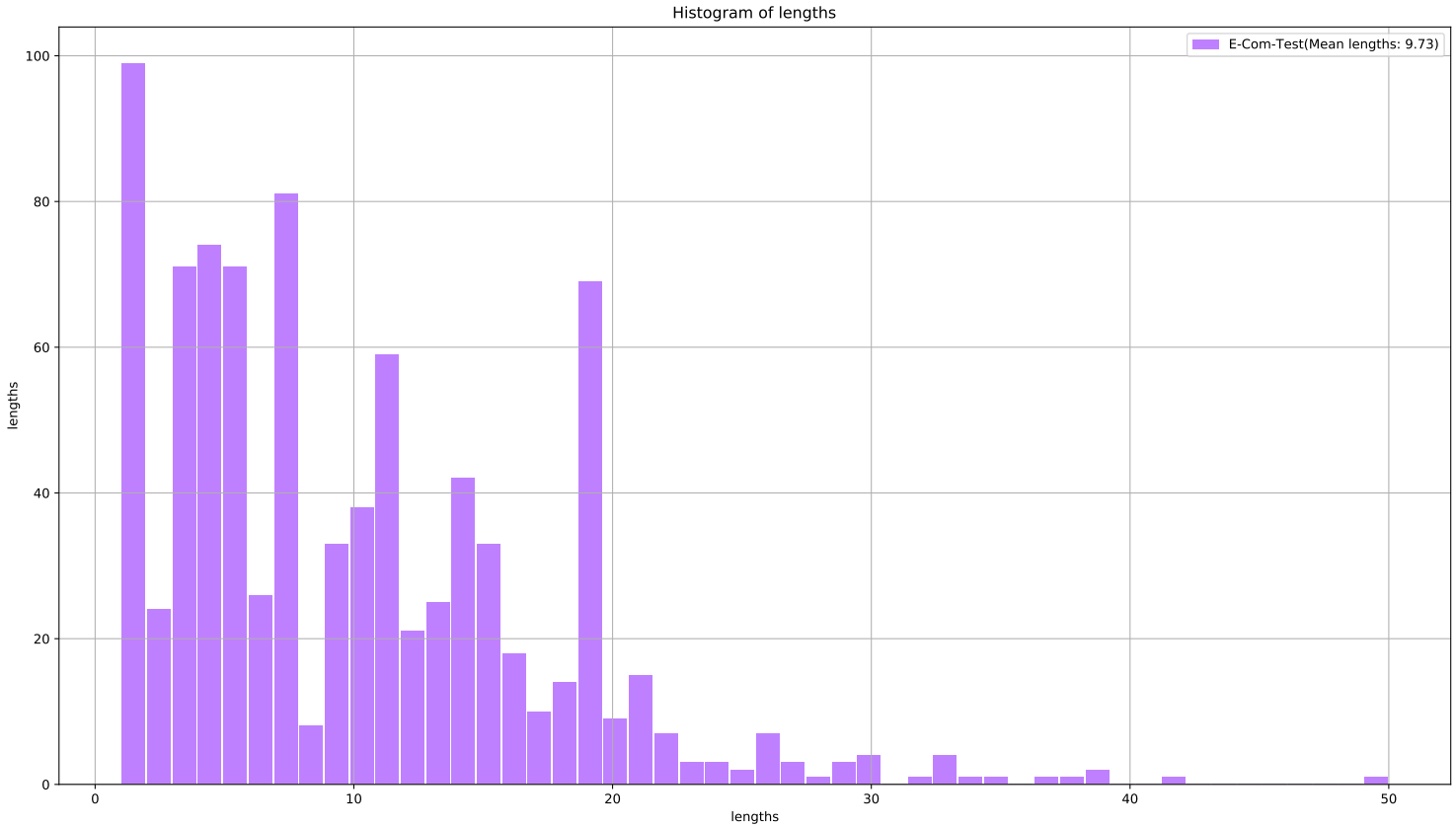
Corpora

Corpus		$\|S\|$	$\|W\|$	$\|V\|$	Avg. length
1 Billion Words	EN	30.3M	800M	800k	26.4
COMMON	EN ES	1.5M	30M 31M	456k 522k	20.0 20.6
dev2013	EN ES	2.7k	48.9k 52.6k	7.5k 9.1k	18.1 19.5
XRCE – Test	EN ES	1.1k	8.4k 10.1k	1.6k 1.7k	7.6 9.2
IT – Test	EN ES	857	15.6k 17.4k	2.1k 2.4k	18.2 20.3
E-Com – Test	EN ES	886	7.3k 8.6k	874 973	8.2 9.7
UE – Test	EN ES	800	22.7k 22.0k	3.9k 4.4k	28.4 27.5
Khreshmoi – Test	EN ES	1000	21.4k 23.9k	4.7 k 4.6k	21.4 23.9

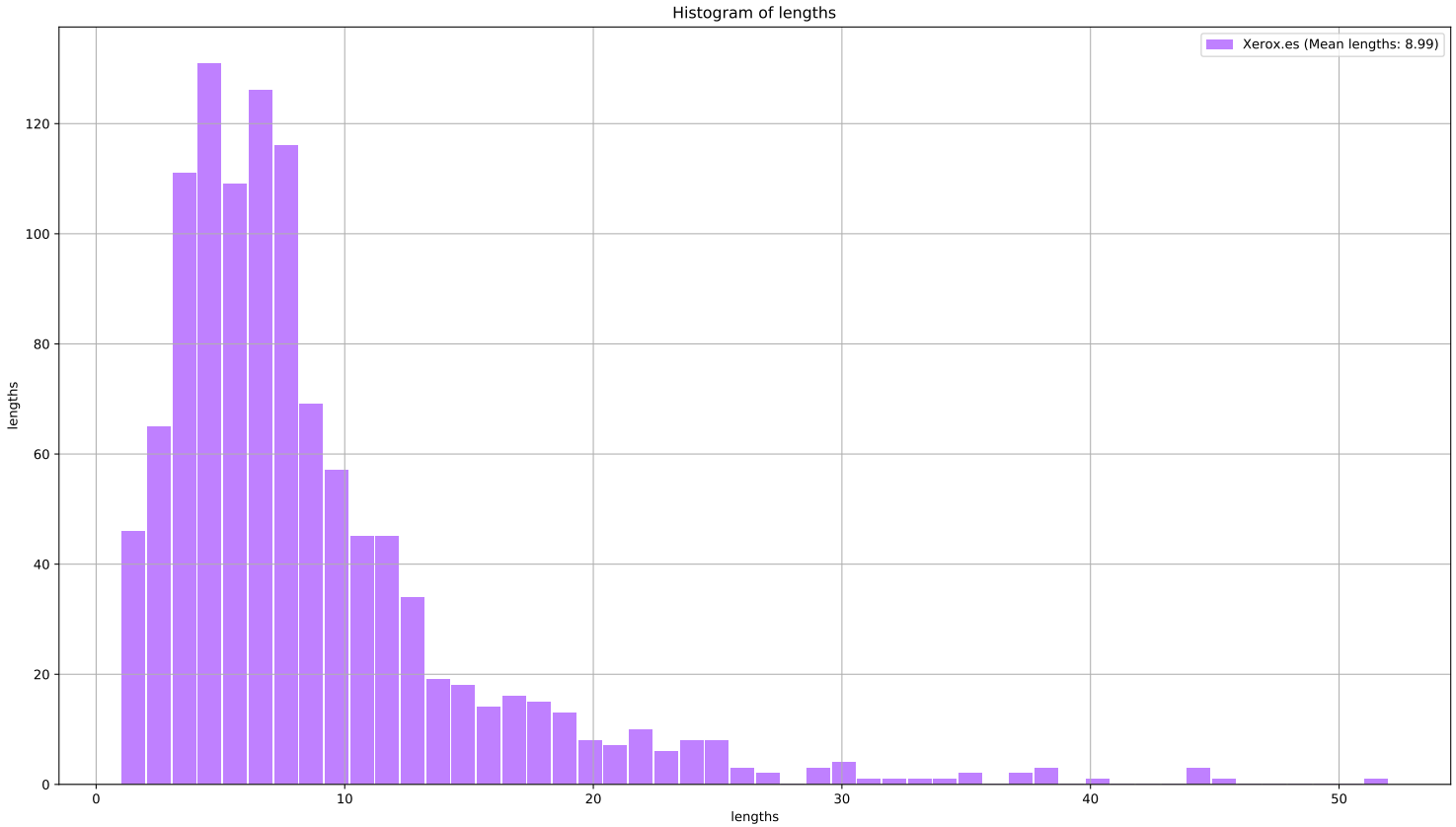
Corpora

Corpus		$\|S\|$	$\|W\|$	$\|V\|$	Avg. length
1 Billion Words	EN	30.3M	800M	800k	26.4
COMMON	EN ES	1.5M	30M 31M	456k 522k	20.0 20.6
dev2013	EN ES	2.7k	48.9k 52.6k	7.5k 9.1k	18.1 19.5
XRCE – Test	EN ES	1.1k	8.4k 10.1k	1.6k 1.7k	**7.6** **9.2**
IT – Test	EN ES	857	15.6k 17.4k	2.1k 2.4k	18.2 20.3
E-Com – Test	EN ES	886	7.3k 8.6k	874 973	**8.2** **9.7**
UE – Test	EN ES	800	22.7k 22.0k	3.9k 4.4k	28.4 27.5
Khreshmoi – Test	EN ES	1000	21.4k 23.9k	4.7 k 4.6k	21.4 23.9

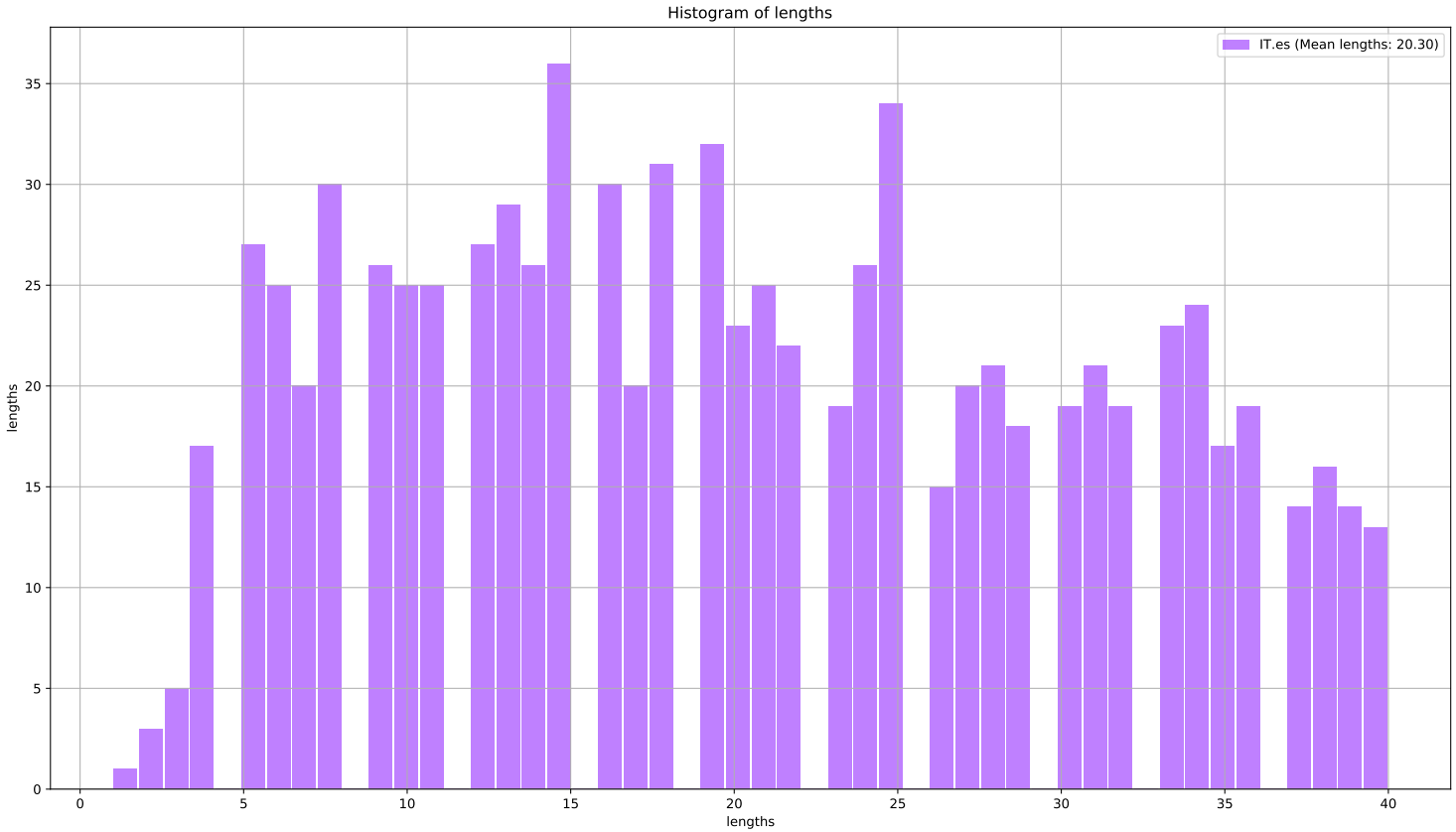
Histogram of lengths: E-Com



Histogram of lengths: Xerox



Histogram of lengths: IT



Results: Xerox

Table 1: Average reference length: 9.2 words.

	XRCE				
System	BLEU	TER	METEOR	$\|W\|$	$\|\bar{W}\|$
Moses	26.2 ± 0.8	59.0 ± 0.8	50.0	$10.3k$	9.1
NMT	20.4 ± 1	94.5 ± 5.1	43.8	$14.4k$	12.8
NMT$^{\Sigma}$	25.5 ± 0.8	76.8 ± 2.0	48.8	$12.7k$	11.3
NMT + Synthetic	27.5 ± 0.8	56.7 ± 0.8	49.0	$9.6k$	8.6
NMT$^{\Sigma}$ + Synthetic	27.3 ± 0.8	56.3 ± 0.8	48.8	$9.4k$	8.4

Results: IT

Table 2: Average reference length: 20.3 words.

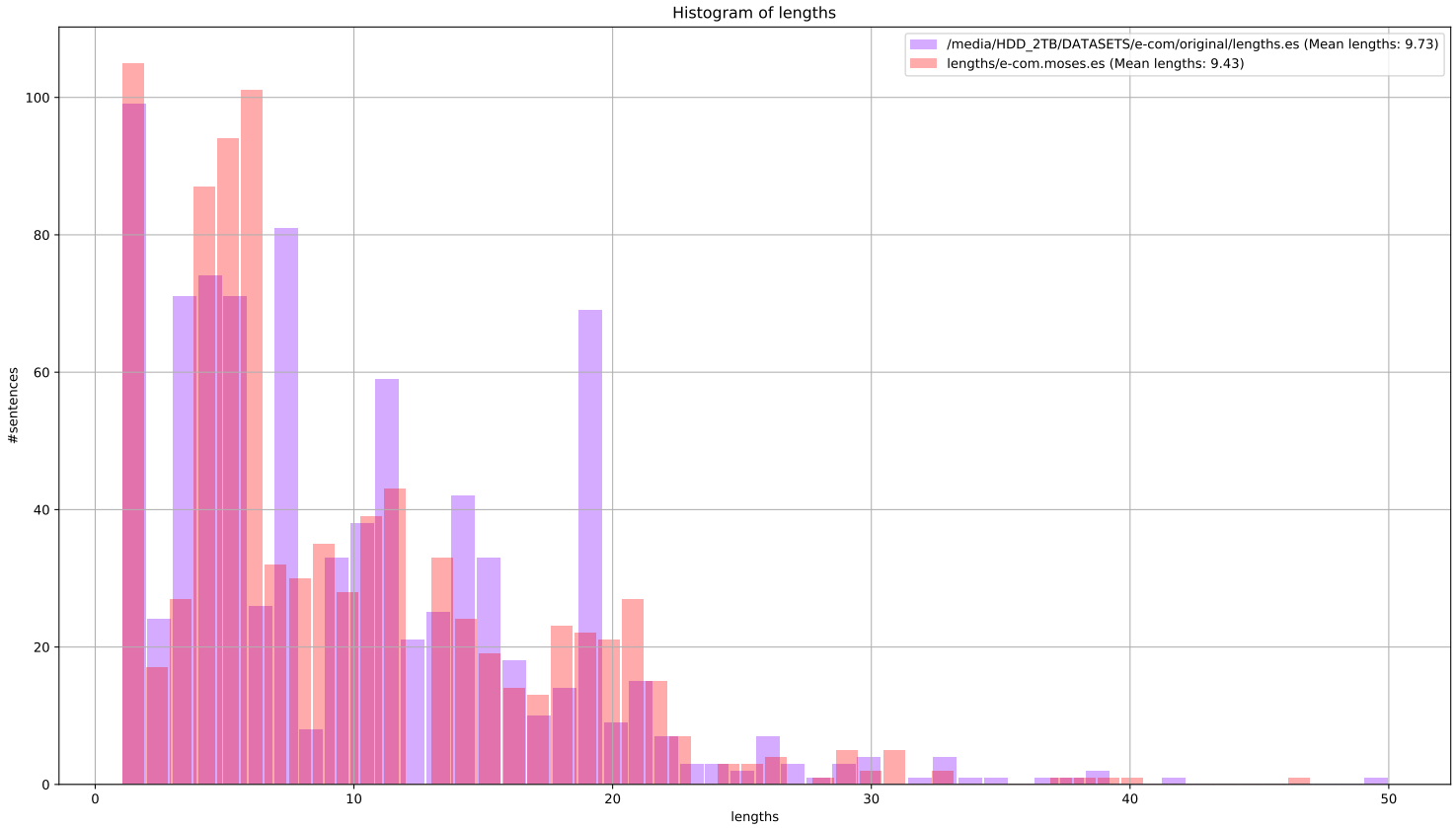
	IT				
System	BLEU	TER	METEOR	$\lvert W \rvert$	$\lvert \bar{W} \rvert$
Moses	33.4 ± 0.6	45.6 ± 0.6	58.1 ± 0.5	17.5k	20.4
NMT	29.0 ± 0.8	53.5 ± 0.8	51.5 ± 0.8	13.1k	15.3
NMT$^{\Sigma}$	31.4 ± 0.8	51.2 ± 0.8	53.9 ± 0.8	13k	15.3
NMT + Synthetic	34.1 ± 0.7	45.7 ± 0.7	57.9 ± 0.6	15.3	17.8
NMT$^{\Sigma}$ + Synthetic	33.8 ± 0.7	46.3 ± 0.7	57.5 ± 0.7	15.5	18.1

Results

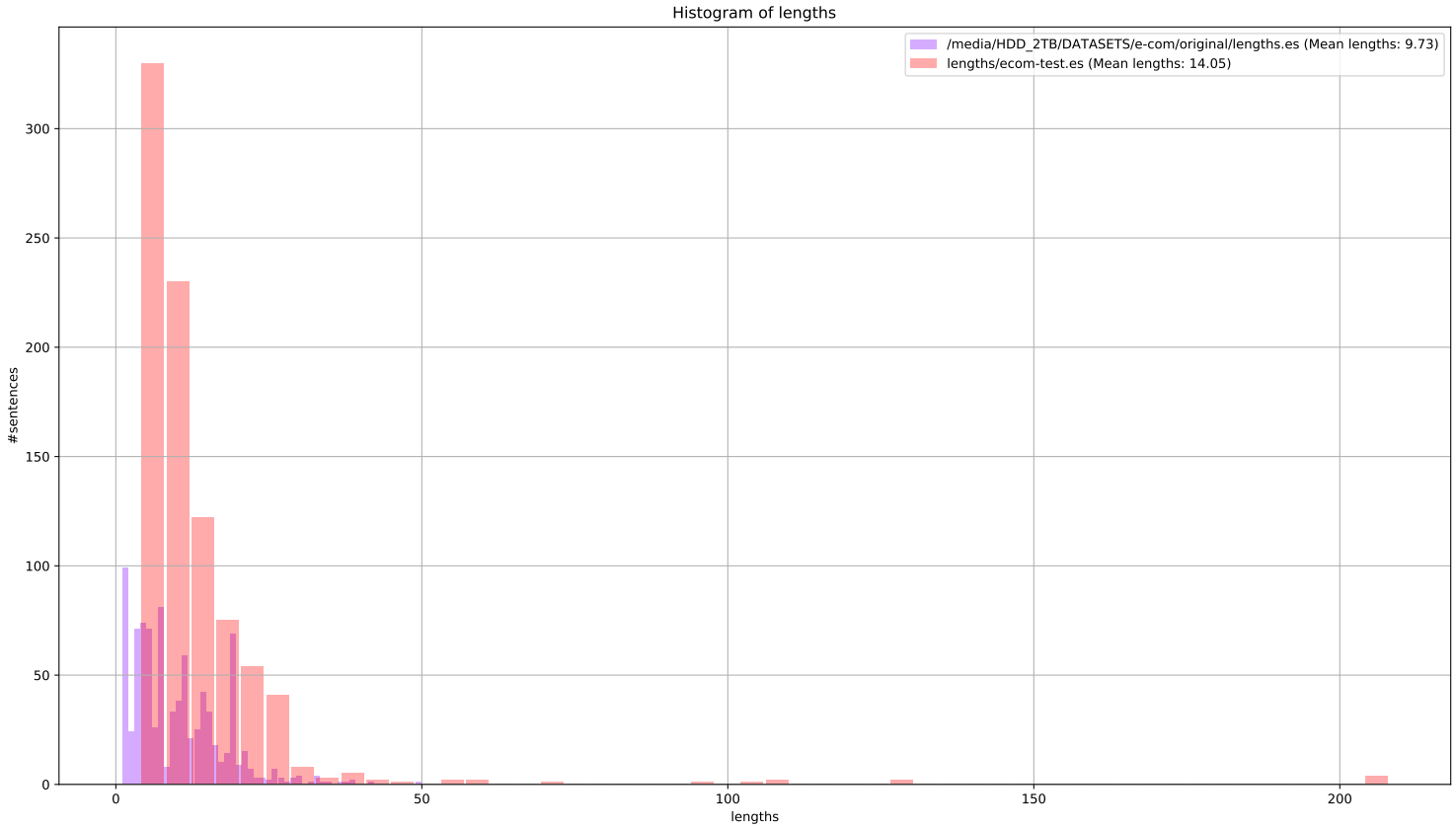
Table 3: Average reference length: 9.7 words.

	E-Com				
System	BLEU	TER	METEOR	$\lvert W \rvert$	$\lvert \bar{W} \rvert$
Moses	21.1 ± 0.8	56.7 ± 0.7	48.7 ± 0.6	8.4k	9.4
NMT	16.9 ± 1.0	104.7 ± 6.3	36.8 ± 1.2	12.5k	14.1
NMT$^{\Sigma}$	23.0 ± 1.0	80.8 ± 2.9	44.0 ± 1.0	10.6k	12.0
NMT + Synthetic	25.5 ± 1.0	59.1 ± 1.0	44.8 ± 0.8	7.7k	8.7
NMT$^{\Sigma}$ + Synthetic	25.8 ± 1.0	61.1 ± 2.6	45.1 ± 1.0	7.7k	8.7

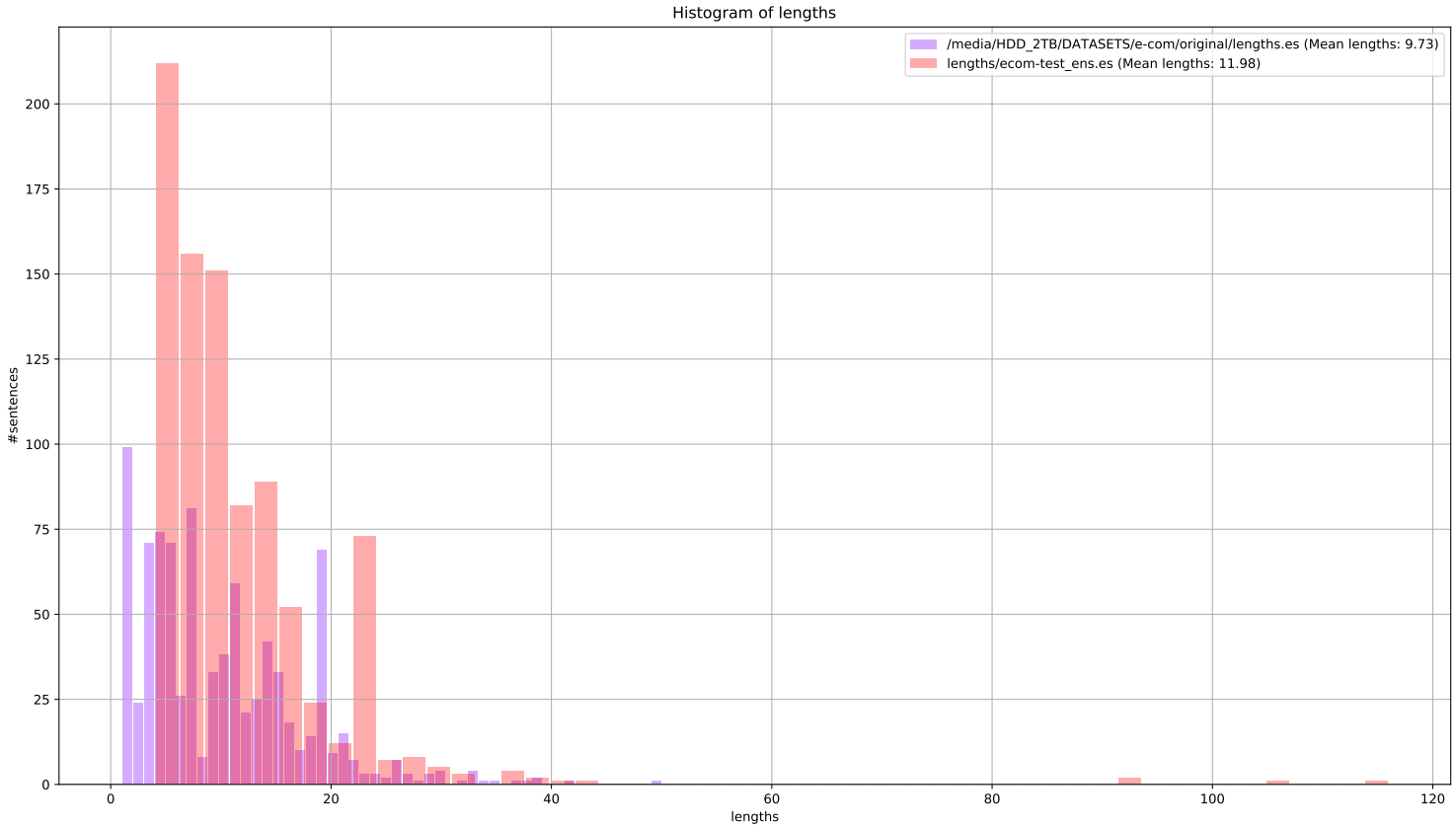
Histogram of lengths: E-Com Moses



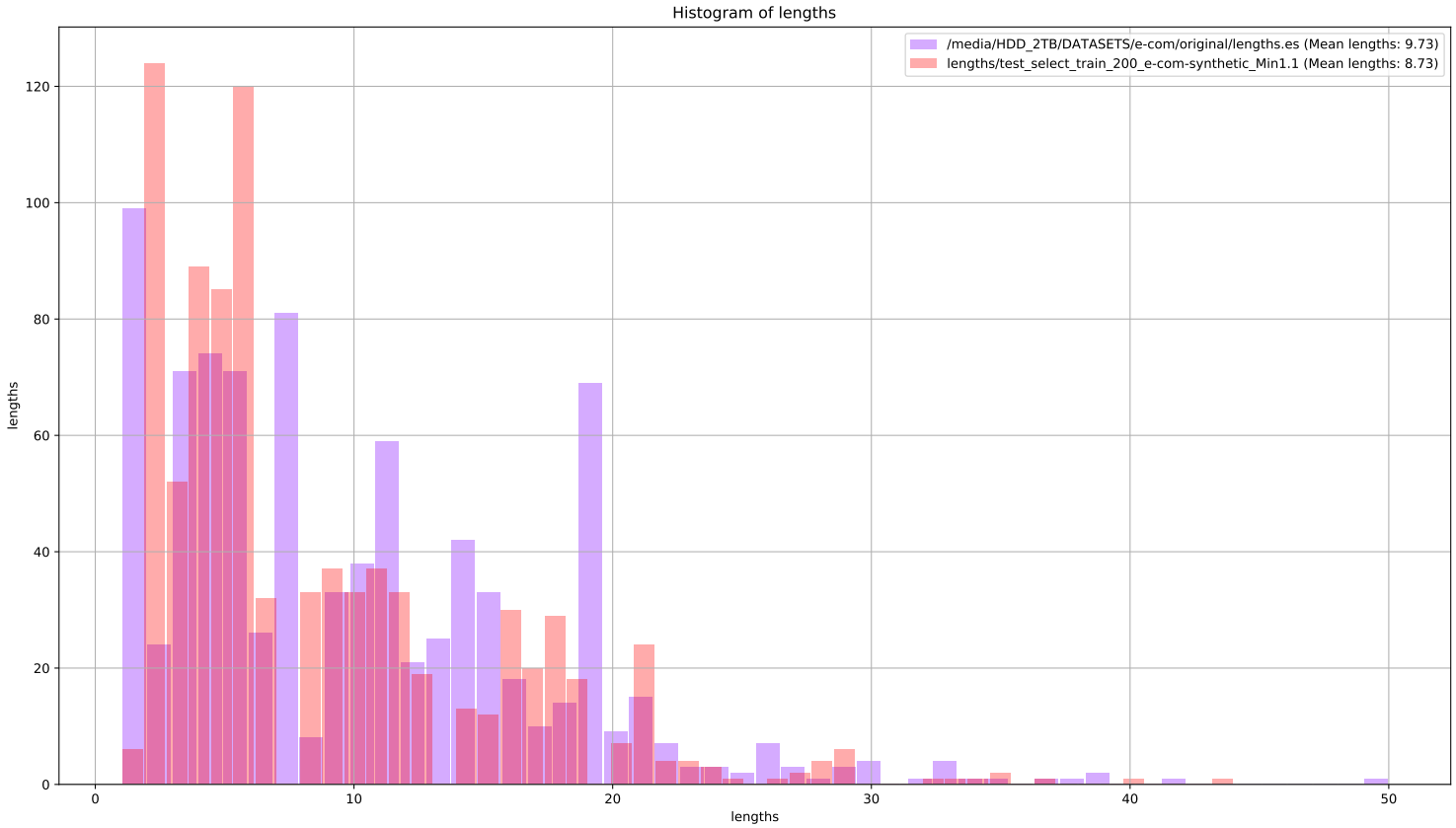
Histogram of lengths: E-Com NMT



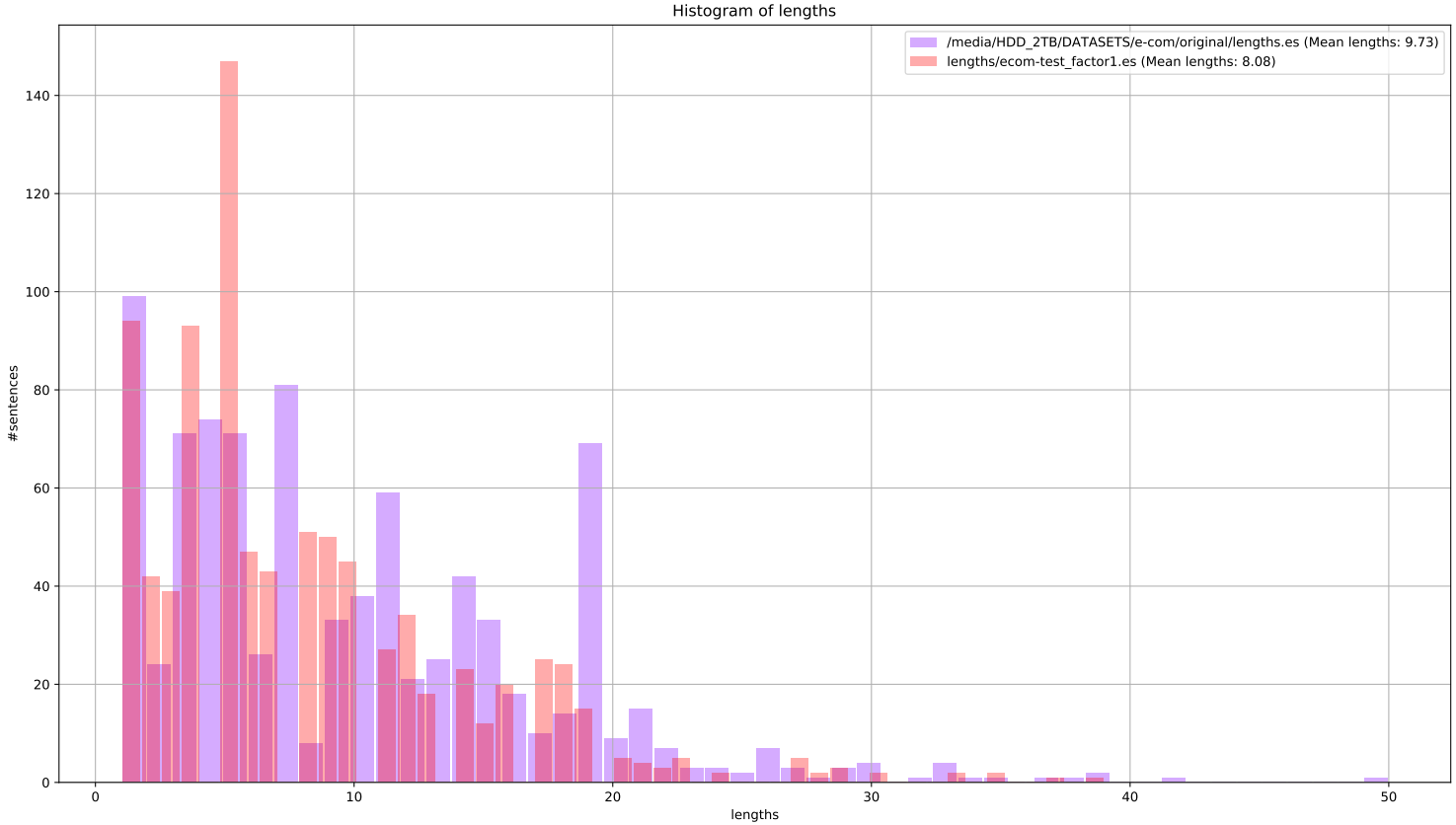
Histogram of lengths: E-Com NMT$^{\Sigma}$



Histogram of lengths: E-Com Synthetic NMT



E-Com NMT – Limiting maximum output length to $|x|$



More results

https://drive.google.com/open?id=1AEPNvOnlkP7IZ8XMqqaLdGpFoul2Bgd1X8bhf07u

Conclusions

- To use a single automatic metric for evaluating machine translation is risky.
- With real datasets, we can obtain strange behaviors of the metrics.
- When applying NMT to tasks with different features than the training data, we should control the length of the output sentences.
- This can be achieved with:
  - Heuristics.
  - Adapting with an in-domain corpus.