LREC-COLING 2024

COMET for Low-Resource Machine Translation Evaluation

A Case Study of English→Maltese and Spanish→Basque

Júlia Falcão[1,2] Claudia Borg[1] Nora Aranberri[2] Kurt Abela[1]

1 L-Università ta' Malta

2 eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

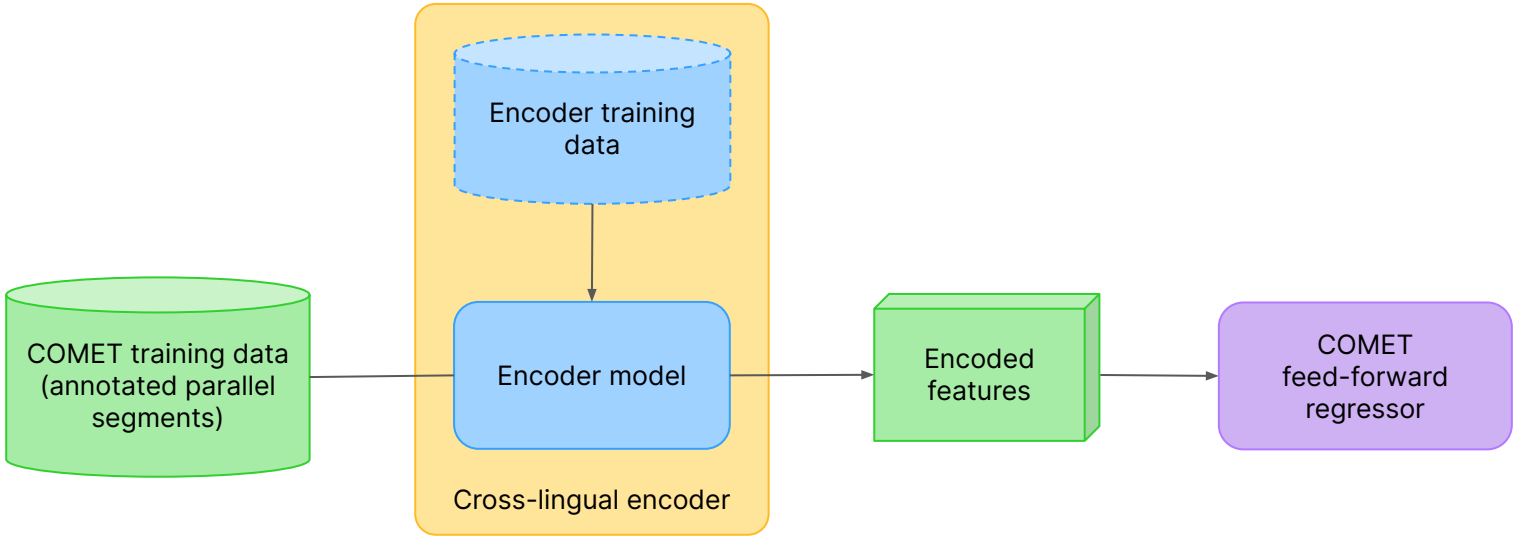
With the support of the Erasmus+ Programme of the European Union

Motivation

- Evaluation has always been a challenge in machine translation
- Automatic metrics based on lexical overlap (e.g. **BLEU**, WER, CHRF) are widely used because they're practical, but research shows they correlate poorly with human judgements
- Newer approach: **neural metrics** that directly optimize for correlation with human judgements
    - COMET being the most well-known so far
    - But neural metrics have limited language support

COMET Framework

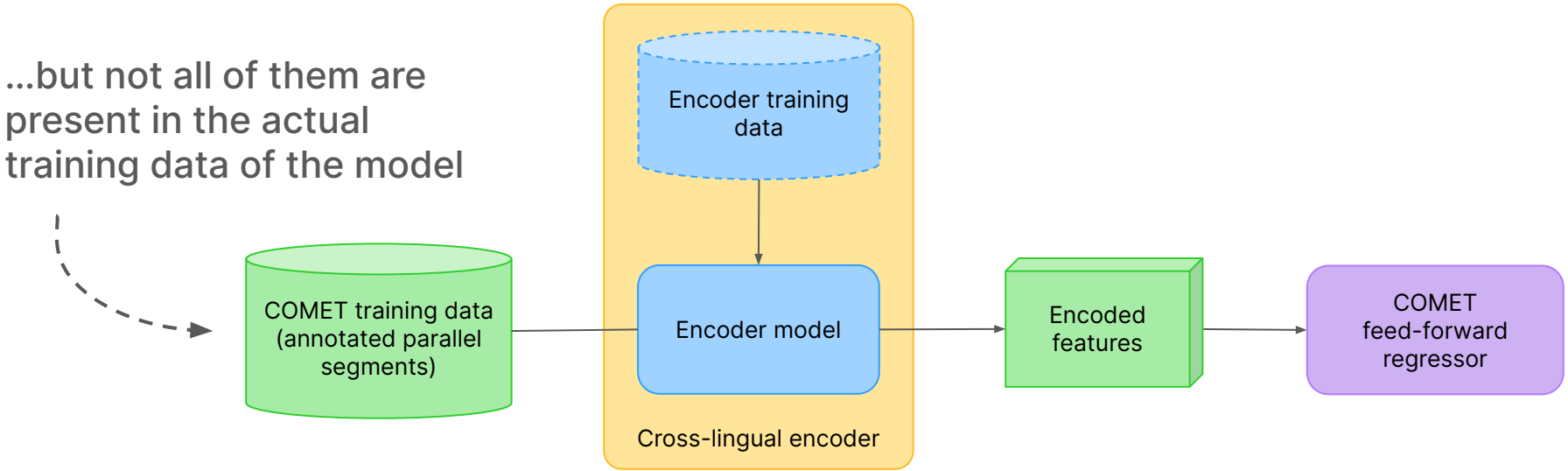
- Neural models trained to predict human scores of quality
- Trained on parallel data annotated with human judgements, embedded by a cross-lingual encoder



COMET language support

In theory, COMET supports all languages supported by the encoder...

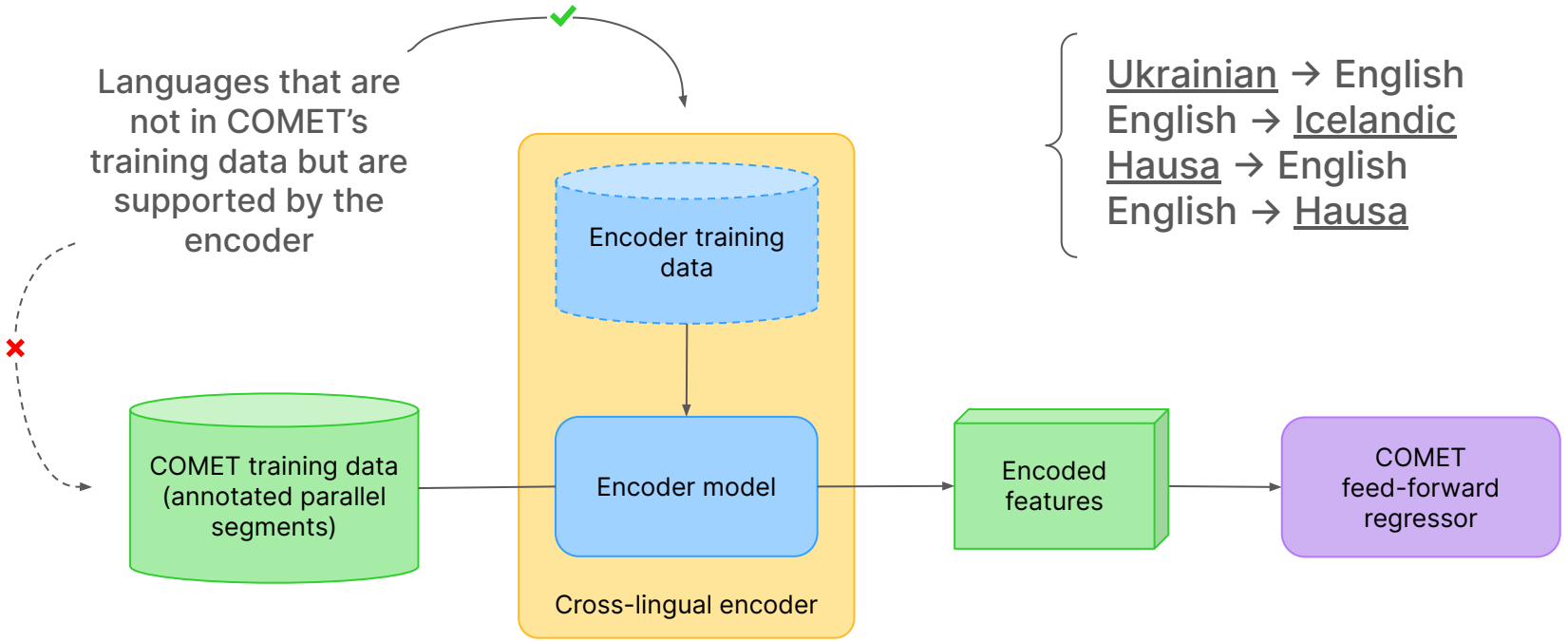
...but not all of them are present in the actual training data of the model



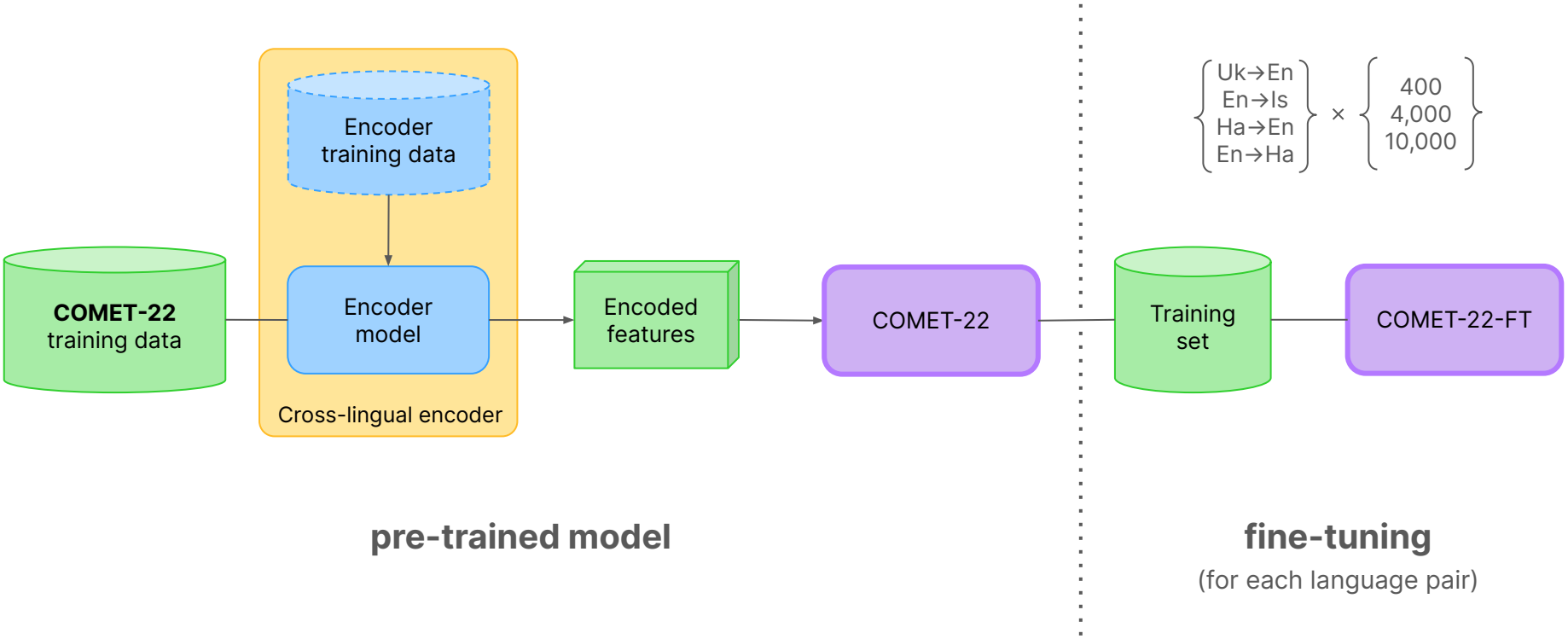
Problem statement

- COMET has shown good performance in meta-evaluations
- But it is mostly tested on the same language pairs it was trained on, the majority being high-resource languages
- Under-resourced languages are at risk of being left out of such advancements
- **How well can COMET models evaluate other languages besides the ones it was trained on?**
- **How could we extend COMET models to include under-resourced languages?**

Preliminary experiments



Preliminary experiments



Results: correlation coefficients

	Uk→En		En→Is		Ha→En		En→Ha	
	τ	**ρ**	**τ**	**ρ**	**τ**	**ρ**	**τ**	**ρ**
COMET-22 (baseline)	0.017	0.025	**0.423**	**0.589**	0.110	0.159	0.145	0.190
Small (400 samples)	0.025	0.038	0.426	0.591	0.106	0.152	0.112	0.147
Medium (4,000 samples)	0.087	0.123	0.476	0.657	0.111	0.156	0.194	0.255
Large (10,000 samples)	0.099	0.139	**0.488**	**0.673**	0.082	0.114	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

τ: Kendall's Tau, **ρ**: Spearman correlation

* Scores in red were deemed statistically insignificant

Results: correlation coefficients

	Uk→En		En→Is		Ha→En		En→Ha	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
COMET-22 (baseline)	0.017	0.025	**0.423**	**0.589**	0.110	0.159	0.145	0.190
Small (400 samples)	0.025	0.038	0.426	0.591	0.106	0.152	0.112	0.147
Medium (4,000 samples)	0.087	0.123	0.476			6	0.194	0.255
Large (10,000 samples)	0.099	0.139	**0.488**			4	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

Low scores from COMET-22, except for En→Is, the pair where the languages are most closely related

Results: correlation coefficients

	Uk→En		En→Is		Ha→En		En→Ha	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
COMET-22 (baseline)	0.017	0.025	0.423	0.589	0.110	0.159	0.145	0.190
Small (400 samples)	0.025	0.038	0.426			0.152	0.112	0.147
Medium (4,000 samples)	0.087	0.123	0.476			0.156	0.194	0.255
Large (10,000 samples)	0.099	0.139	0.488	0.673	0.082	0.114	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

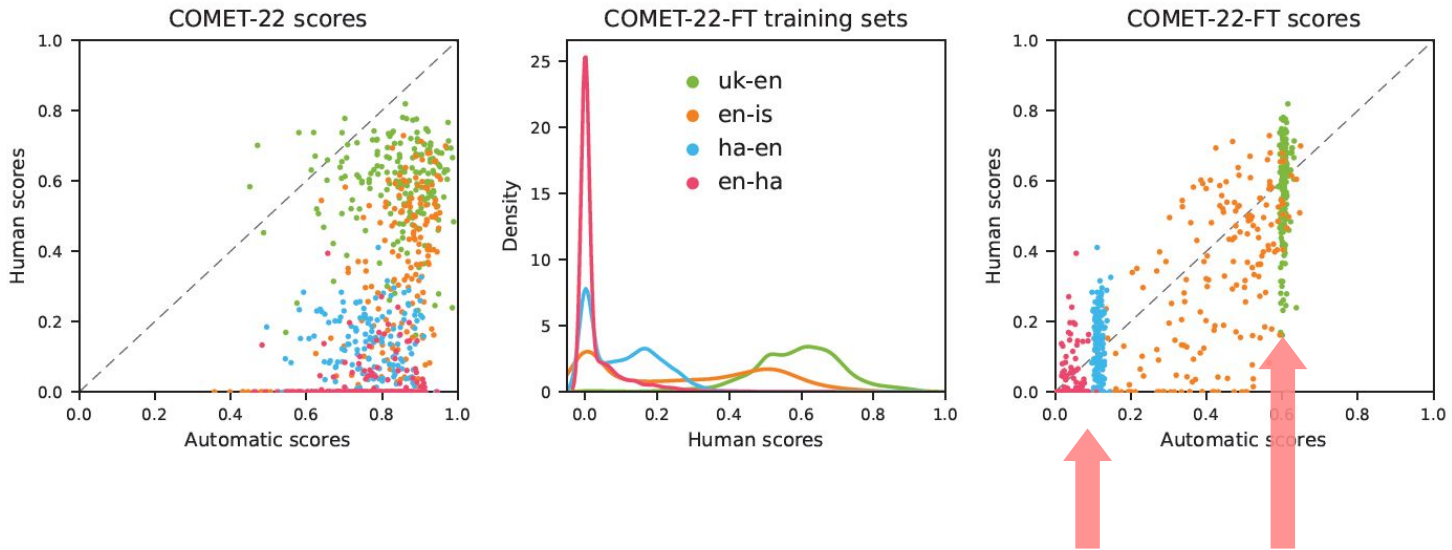
Statistically insignificant correlations for Uk→En, only pair with a different data domain, but it had the largest improvement

Results: correlation coefficients

	Uk→En		En→Is		**Ha→En**		En→Ha	
	τ	ρ	τ	ρ	**τ**	**ρ**	τ	ρ
COMET-22 (baseline)	0.017	0.025	**0.423**	**0.589**	0.110	0.159	0.145	0.190
Small (400 samples)	0.025			0.591	0.106	0.152	0.112	0.147
Medium (4,000 samples)	0.087			0.657	0.111	0.156	0.194	0.255
Large (10,000 samples)	0.099	0.139	**0.488**	**0.673**	0.082	0.114	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

The only language pair that did not improve with fine-tuning was Ha→En, possibly due to overfitting

Analysis: quality score distributions



Models only produced scores within **narrow ranges** that match the highest concentrations of scores in their training data

Manual evaluation campaign

Online campaign to collect **human judgements of translation quality** for English→Maltese and Spanish→Basque

- Participants: volunteer bilingual speakers, a mix of experts and non-experts
- Implemented on the Appraise evaluation framework
- The participants could complete as many evaluations as they wished

Data: original datasets

Source text	Nowadays air travel is only rarely booked directly through the airline without first searching and comparing prices.
Reference translation	Illum il-ġurnata l-ivvjaġġar bl-ajru rari jiġi bbukkjat direttament permezz tal-linja tal-ajru mingħajr ma l-ewwel isir tiftix u paragunar tal-prezzijiet.

Parallel set of 400 segments and reference translations for each language pair, manually selected

Data: translation hypotheses

Source text	Nowadays air travel is only rarely booked directly through the airline without first searching and comparing prices.
Reference translation	Illum il-ġurnata l-ivvjaġġar bl-ajru rari jiġi bbukkjat direttament permezz tal-linja tal-ajru mingħajr ma l-ewwel isir tiftix u paragunar tal-prezzijiet.
MT output	Illum il-ġurnata l-ivvjaġġar bl-ajru huwa rari biss ibbukkjat direttament permezz tal-linja tal-ajru mingħajr ma l-ewwel wieħed ifittex u jqabbel il-prezzijiet.

Automatic translations from **3 systems**

- one **proprietary** system
- one **open-source** model
- one new, **in-house** model

Data: damaged outputs

Source text	Nowadays air travel is only rarely booked directly through the airline without first searching and comparing prices.
Reference translation	Illum il-ġurnata l-ivvjaġġar bl-ajru rari jiġi bbukkjat direttament permezz tal-linja tal-ajru mingħajr ma l-ewwel isir tiftix u paragunar tal-prezzijiet.
MT output	Illum il-ġurnata l-ivvjaġġar bl-ajru huwa rari biss ibbukkjat direttament permezz tal-linja tal-ajru mingħajr ma l-ewwel wieħed ifittex u jqabbel il-prezzijiet.
Another randomly selected reference	Dan is-servizz huwa b'xejn, u b'mod faċli, jista' jkollok aċċess għall-aħbarijiet riċenti, fatti u figuri, dokumenti legali u għadd kbir ta' informazzjoni prattika.
Damaged output	Illum il-ġurnata l-ivvjaġġar bl-ajru huwa rari biss ibbukkjat direttament permezz tal-linja tal-ajru mingħajr jkollok aċċess għall-aħbarijiet riċenti, fatti jqabbel il-prezzijiet.

Quality control items

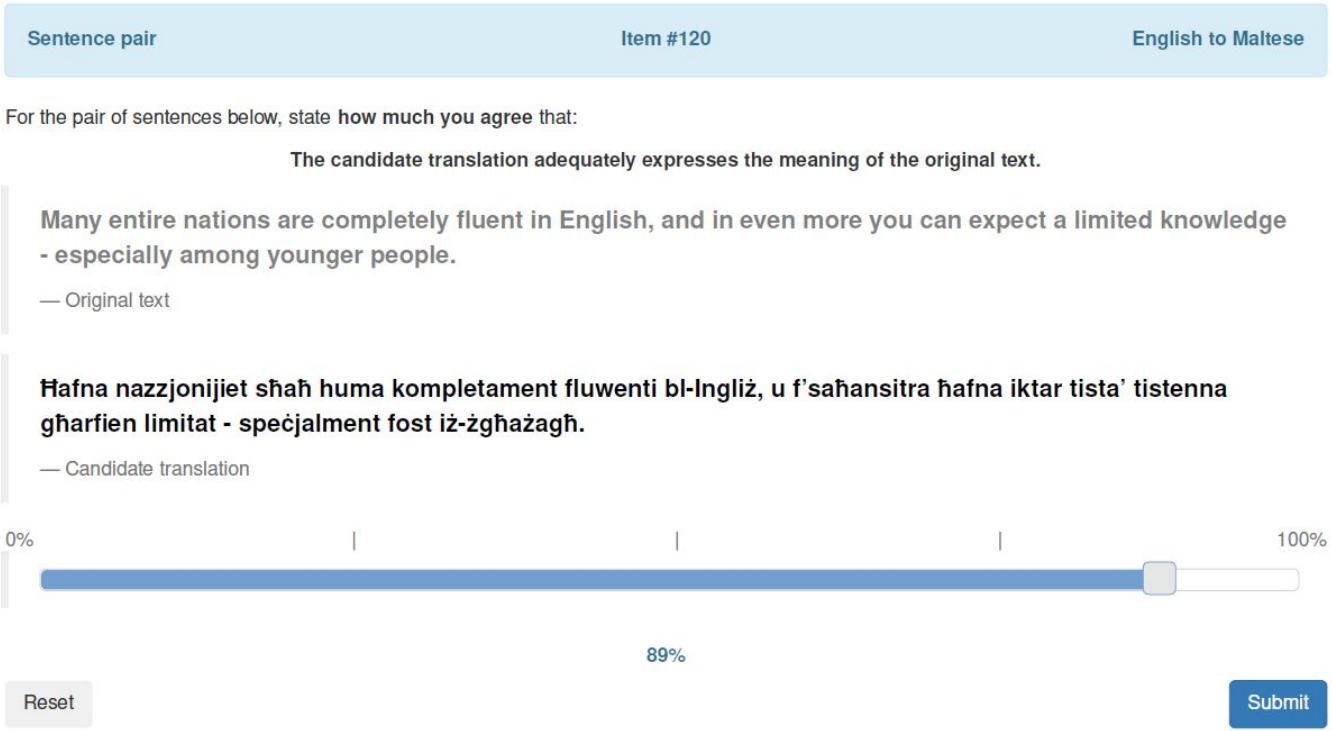
Source text	Nowadays air travel is only rarely booked directly through the airline without first searching and comparing prices.
Reference translation	Illum il-ġurn… permezz ta… …unar tal-prezzijiet.
MT output	Illum il-ġurn… direttament… …ned ifittex u jqabbel il-prezzijiet.
Another randomly selected reference	Dan is-servizz huwa b'xejn, u b'mod faċli, jista' jkollok aċċess għall-aħbarijiet riċenti, fatti u figuri, dokumenti legali u għadd kbir ta' informazzjoni prattika.
Damaged output	Illum il-ġurn… direttament… għall-aħbarijiet riċenti, fatti jqabbel il-prezzijiet.

expected to receive high score

evaluation target

expected to receive low score

The task: direct assessment (DA)



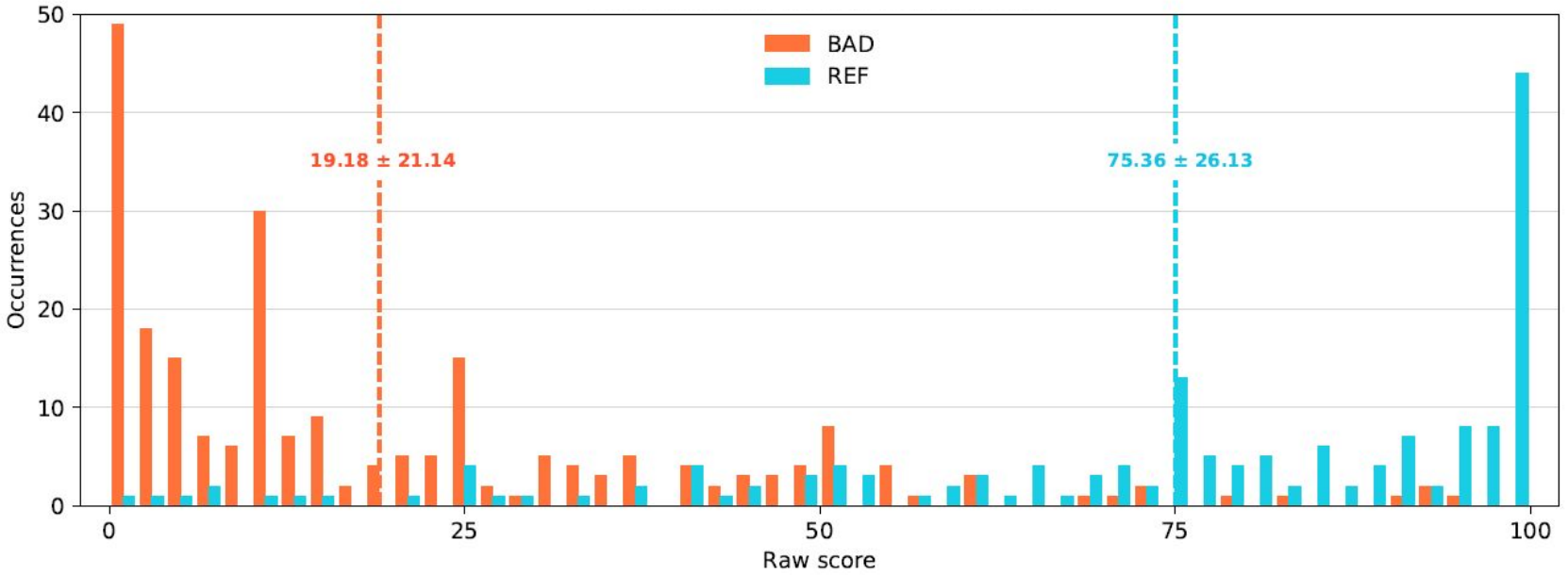
Turnout

	En→Mt	Es→Eu
Total evaluations	992	1215
↳ MT outputs	811	996
↳ Damaged outputs	101	133
↳ Reference texts	80	86
Total participants	41	44
Avg. evaluations per user	24	27
Quality control failures	8	11
↳ Damaged outputs	4	2
↳ Reference texts	4	9
“Unreliable” participants	5	8
Discarded evaluations	183	361
Remaining evaluations	628	635

Raw evaluation results

Quality control filtering

Quality control item scores



Results: out-of-the-box evaluations

	Systems	Automatic metrics				Human evaluation	
		COMET-22	BLEU	TER	CHRF	Raw	Z-scores
En→Mt	GT	0.74 #1	44 #1	39 #1	74 #1	82 #1	0.59 #1
	NLLB	0.69 #2	25 #2	64 #3	63 #2	65 #2	0.12 #2
	UM-IWSLT	0.69 #3	24 #3	61 #2	59 #3	49 #3	-0.43 #3
Es→Eu	Itzuli	0.84 #2	15 #3	79 #3	54 #3	83 #1	0.43 #1
	NLLB	0.83 #3	27 #1	69 #1	57 #1	64 #3	-0.17 #3
	UPV-CMBT	0.84 #1	16 #2	79 #2	54 #2	82 #2	0.36 #2

Results: out-of-the-box evaluations

	Systems	Automatic metrics				Human evaluation	
		COMET-22	BLEU	TER	CHRF	Raw	Z-scores
En→Mt	GT	0.74 #1	44 #1	39 #1	74 #1	82 #1	0.59 #1
	NLLB	0.69 #2	25 #2	64 #3	63 #2	65 #2	0.12 #2
	UM-IWSLT	0.69 #3	24 #3	61 #2	59 #3	49 #3	-0.43 #3
Es→Eu	Itzuli	0.84 #2	15 #3	79 #3	54 #3	83 #1	0.43 #1
	NLLB	0.83 #3	27 #1	69 #1	57 #1	64 #3	-0.17 #3
	UPV-CMBT	0.84 #1	16 #2	79 #2	54 #2	82 #2	0.36 #2

All metrics and human evaluation agreed on the **ranking** of systems for En→Mt:

1. Google Translate
2. NLLB
3. UM-IWSLT

Results: out-of-the-box evaluations

	Systems	Automatic metrics				Human evaluation	
		COMET-22	BLEU	TER	CHRF	Raw	Z-scores
En→Mt	GT	0.74 #1	44 #1	39 #1	74 #1	82 #1	0.59 #1
	NLLB	0.69 #2	25 #2	64 #3	63 #2	65 #2	0.12 #2
	UM-IWSLT	0.69 #3	24 #3	61 #2	59 #3	49 #3	-0.43 #3
Es→Eu	Itzuli	0.84 #2	15 #3	79 #3	54 #3	83 #1	0.43 #1
	NLLB	0.83 #3	27 #1	69 #1	57 #1	64 #3	-0.17 #3
	UPV-CMBT	0.84 #1	16 #2	79 #2	54 #2	82 #2	0.36 #2

For Es→Eu, COMET-22 and humans deemed **Itzuli** and **UPV-CMBT** to be the best, and of almost equal quality...

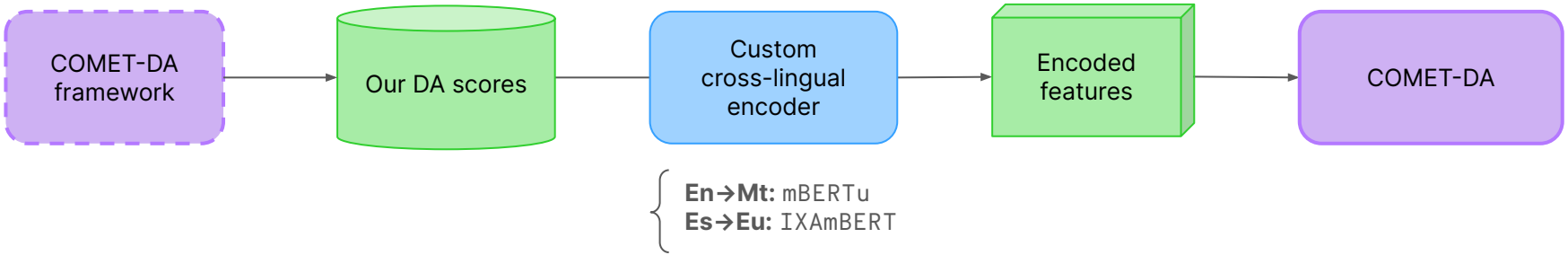
...but lexical overlap metrics rated **NLLB** as the best, by a significant margin.

Results: out-of-the-box evaluations

	Systems	Automatic metrics				Human evaluation	
		COMET-22	BLEU	TER	CHRF	Raw	Z-scores
En→Mt	GT	0.74 #1	44 #1	39 #1	74 #1	82 #1	0.59 #1
	NLLB	0.69 #2	25 #2	64 #3	63 #2	65 #2	0.12 #2
	UM-IWSLT	0.69 #3	24 #3	61 #2	59 #3	49 #3	-0.43 #3
Es→Eu	Itzuli	0.84 #2	15 #3	79 #3	54 #3	83 #1	0.43 #1
	NLLB	0.83 #3	27 #1	69 #1	57 #1	64 #3	-0.17 #3
	UPV-CMBT	0.84 #1	16 #2	79 #2	54 #2	82 #2	0.36 #2

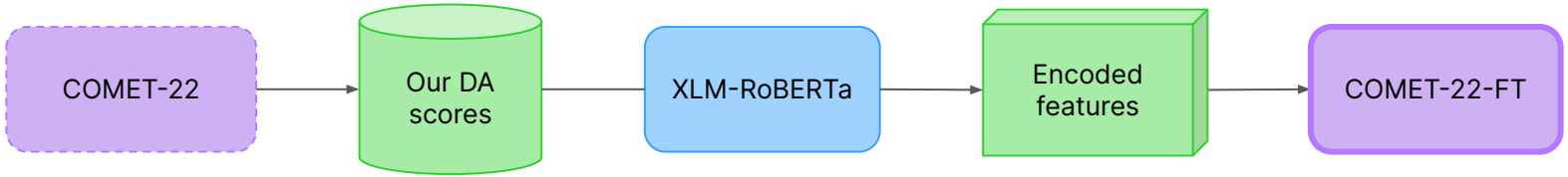
COMET-22 rated all systems **very closely**, as opposed to larger deltas from other metrics and from humans.

Improvement strategy #1: training from scratch



- New COMET-DA model trained from scratch on our evaluation results
- **Pros:** switching the encoder
- **Cons:** small training set of our evaluation results

Improvement strategy #2: fine-tuning



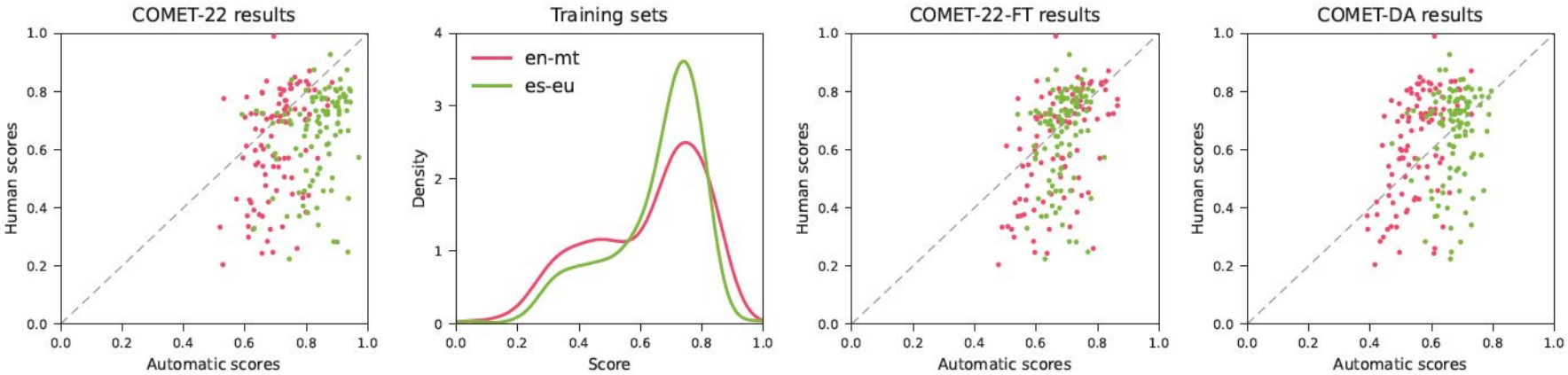
- Using our evaluation results to further train COMET-22
- **Pros:** COMET-22 has the potential to generalize
- **Cons:** XLM-R vocabulary issue for Maltese

Results: correlation scores

- **COMET-22-FT** obtained the highest correlations, showing improvement over the base model
  - Similar findings to previous research on the potential of fine-tuning COMET models
- **COMET-DA** models performed very differently
  - High correlations for En→Mt, but statistically insignificant values for Es→Eu

	Model	τ	ρ
En→Mt	**COMET-22**	0.29	0.42
	COMET-DA	0.38	0.53
	COMET-22-FT	**0.39**	**0.54**
Es→Eu	**COMET-22**	0.22	0.33
	COMET-DA	0.12	0.17
	COMET-22-FT	**0.25**	**0.35**

Results: quality score distributions



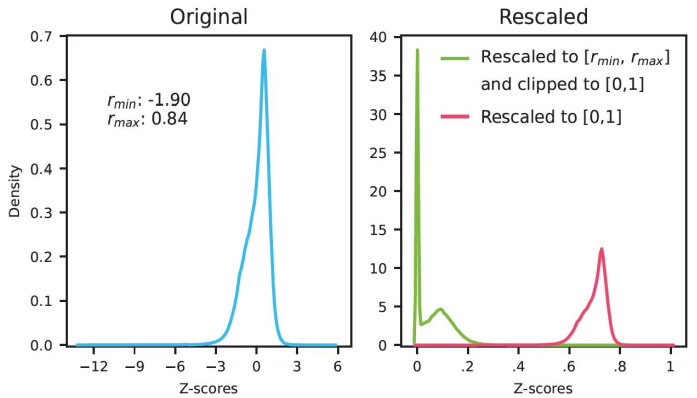
En→Mt: slightly more balanced training set ⇒ wider variety of produced scores (0.4-0.8)
Es→Eu: less balanced data, mostly around 0.7 ⇒ less variety of quality scores (0.6-0.8)

Distributions appear influenced by the distributions of scores in the training sets

Results: low-scoring test set

- COMET-22 performed much worse on a test set with only scores ≤ 0.6
- Higher correlation on the regular test set might have been out of “luck”, i.e. influenced by its training data distribution

	En→Mt		Es→Eu	
Test set	**τ**	**ρ**	**τ**	**ρ**
Regular	0.29	0.42	0.22	0.33
Low-scoring	0.10	0.14	-0.01	-0.01



Conclusions

- **Fine-tuning can improve the performance of COMET for evaluating language pairs that are not in its training data**
  - Improvement of 0.08 in Kendall's Tau for Uk→En with 10K samples
- **Scores generated by COMET models appeared heavily influenced by the distribution of scores in their training data**
  - COMET-22-FT and COMET-DA mostly produced scores between 0.6-0.8 after being trained on datasets concentrated in this range
  - COMET-22 itself was trained on unbalanced data and performs worse on a low-scoring test set

Main contributions & future work

- **First datasets of human evaluations** for English→Maltese and Spanish→Basque translations
    - Released under an open license
- First **analysis on the usability of COMET** for Maltese and Basque
- Experiment results based on a small-scale evaluation campaign
    - Future works could replicate with larger datasets, comparing different approaches, such as other COMET architectures or other human evaluation methods

LREC-COLING 2024

COMET for Low-Resource Machine Translation Evaluation

A Case Study of English→Maltese and Spanish→Basque

Júlia Falcão[1,2] Claudia Borg[1] Nora Aranberri[2] Kurt Abela[1]

1 L-Università ta' Malta

2 eman ta zabal zazu
Universidad del País Vasco Euskal Herriko Unibertsitatea

With the support of the Erasmus+ Programme of the European Union