Form and Meaning in Intrinsic Multilingual Evaluations

EACL 2026

Wessel Poelman & Miryam de Lhoneux

LAGoM-NLP, Department of Computer Science, KU Leuven

Are certain human languages easier or harder to model?

Are certain human languages easier or harder to model?

How do we evaluate this?

Outline

Common setups

- Multi-parallel datasets

Outline

Common setups

- Multi-parallel datasets
  - FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus

Outline

Common setups

- Multi-parallel datasets
  - FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus
- **One multilingual** model or **multiple monolingual** models

Outline

Common setups

- Multi-parallel datasets
  - FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus
- **One multilingual** model or **multiple monolingual** models
- Intrinsic metrics like perplexity

Outline

Common setups

- Multi-parallel datasets
  - FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus
- **One multilingual** model or **multiple monolingual** models
- Intrinsic metrics like perplexity

Same semantic meaning $\rightarrow$ **fair comparison**

Outline

Common setups

- Multi-parallel datasets
  - FLORES-200, EuroParl, UN Parallel Corpus, Parallel Bible Corpus
- **One multilingual** model or **multiple monolingual** models
- Intrinsic metrics like perplexity

Same semantic meaning $\rightarrow$ **fair comparison**

But intrinsic metrics target **form**, not **meaning**...

Example

Model	Sequence	PPL
A	Sabe jugar al ajedrez	20
B	Do you know how to play chess	22
B	Can you play chess	18

Comparison

Assumptions for fair comparison

1. **Within language** $\rightarrow$ roughly same values ($\approx$ semantics?)

Comparison

Assumptions for fair comparison

1. **Within language** $\rightarrow$ roughly same values ($\approx$ semantics?)
2. **Across languages** $\rightarrow$ consistency (no flipped conclusions)

Metrics

For a sequence S, segmented into w (token), c (character):

$$\text{Negative Log-Likelihood (NLL)} = -\frac{1}{S}\sum_{t=1}^{S} \log P(w_t | w_{<t})$$

$$\text{Perplexity (PPL)} = \exp(\text{NLL})$$

$$\text{Bits per Character (BPC)} = -\frac{1}{S}\sum_{t=1}^{S} \log_2 P(c_t | c_{<t})$$

$$\text{Bits per English Character (BPEC)} = \frac{\text{BPC}_{\text{Target}}}{\text{BPC}_{\text{EN}}}$$

$$\text{Information Parity (IP)} = \frac{\text{BPC}_{\text{EN}}}{\text{BPC}_{\text{Target}}}$$

$$\text{Mean Reciprocal Rank (MRR)} = \frac{1}{S}\sum_{t=1}^{S} \frac{1}{R_t}$$ where R is the ranking over the vocabulary

Metrics

(Long) theoretical outline of the assumptions involved...

Metrics

(Long) theoretical outline of the assumptions involved...

See the paper!

Empirical results!

Setup

Experiments

1. Train multilingual and multiple monolingual models on EuroParl and UN Parallel Corpus
   - "Mainstream" small model architecture, tokenization, hyperparameters, ...

Setup

Experiments

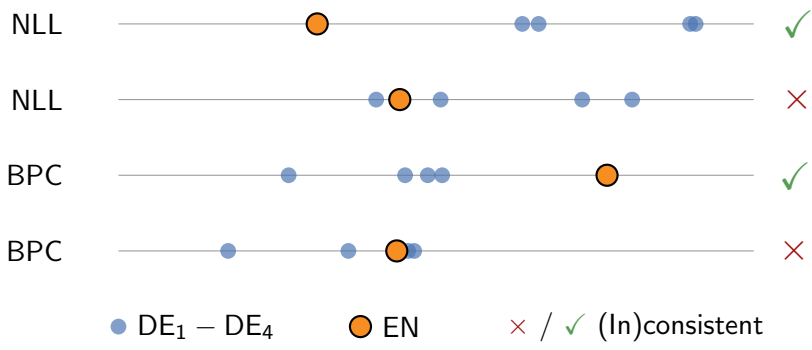
1. Train multilingual and multiple monolingual models on EuroParl and UN Parallel Corpus
   - "Mainstream" small model architecture, tokenization, hyperparameters, ...
2. Evaluate on FLORES-200 and WMT-19 paraphrases
   - One English source with **four** German, human translations

Setup

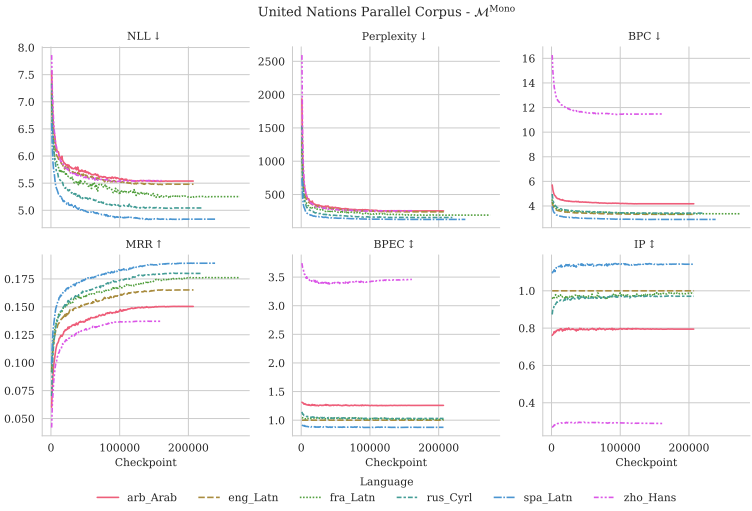
Experiments

1. Train multilingual and multiple monolingual models on EuroParl and UN Parallel Corpus
   - "Mainstream" small model architecture, tokenization, hyperparameters, ...
2. Evaluate on FLORES-200 and WMT-19 paraphrases
   - One English source with **four** German, human translations
3. See what the metrics do:
   - Same values within language
   - Consistency across languages

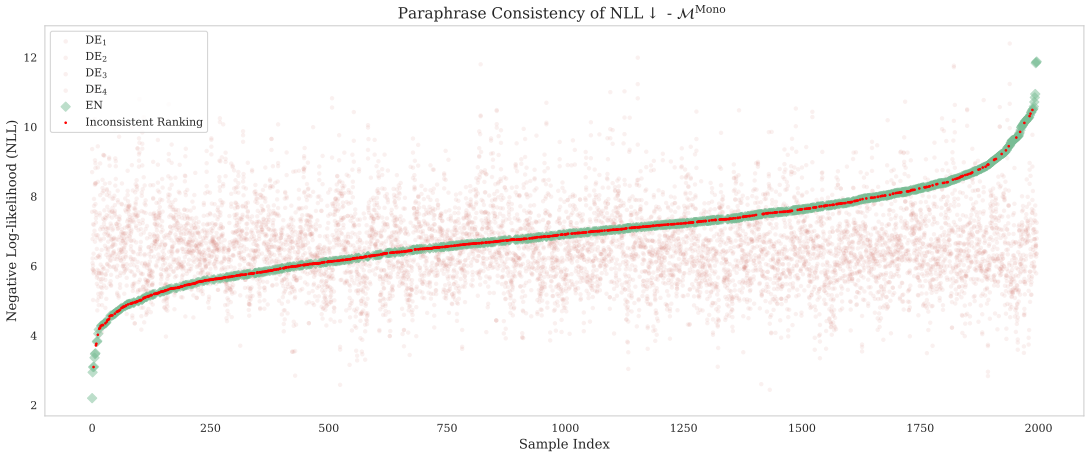
Results



Results


United Nations Parallel Corpus - $\mathcal{M}^{\text{Mono}}$
NLL ↓
Perplexity ↓
BPC ↓
MRR ↑
BPEC ↕
IP ↕
Checkpoint
Language
arb_Arab
eng_Latn
fra_Latn
rus_Cyrl
spa_Latn
zho_Hans

Results


Paraphrase Consistency of NLL ↓ - $\mathcal{M}^{\text{Mono}}$
DE$_1$
DE$_2$
DE$_3$
DE$_4$
EN
Inconsistent Ranking
Negative Log-likelihood (NLL)
Sample Index
2
4
6
8
10
12
0
250
500
750
1000
1250
1500
1750
2000

Results

Metric	$\mathcal{M}^{\text{Mono}}$	$\mathcal{M}^{\text{Multi}}$
NLL	47%	51%
BPC	50%	52%
MRR	50%	51%

Sample inconsistency

Results

Metric	$\mathcal{M}^{\text{Mono}}$	$\mathcal{M}^{\text{Multi}}$
NLL	47%	51%
BPC	50%	52%
MRR	50%	51%

Sample inconsistency

Model	Metric	DE^{F}	EN	DE^{S}
$\mathcal{M}^{\text{Mono}}$	NLL ↓	6.43–6.64 ✓	6.91	5.89–7.23 ×
	BPC ↓	2.09–2.20 ✓	2.27	1.82–2.44 ×
	MRR ↑	19.6–21.2 ✓	18.6	14.6–26.0 ×
$\mathcal{M}^{\text{Multi}}$	NLL ↓	6.81–7.02 ✓	7.05	6.25–7.60 ×
	BPC ↓	2.40–2.51 ✓	2.54	2.13–2.77 ×
	MRR ↑	22.5–23.7 ✓	21.4	17.4–28.7 ×

Split inconsistency

Conclusion

- Consistency in *meaning* does not neutralize differences in **form**.
- Metrics transforming NLL are measuring (and sensitive to) form.
- **Within** languages: not the same values for paraphrases.
- **Across** languages: many inconsistent samples.

Acknowledgements

We thank **Thomas Bauwens** and **Coleman Haley** for discussions and ideas!

WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.

The computational resources and services used were provided by the VSC (Flemish Supercomputer Center),

funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

Thank you! Questions? Feedback?

wessel.poelman@kuleuven.be