Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer

Eleftheria Briakou
ebriakou@cs.umd.edu

Sweta Agrawal
sweagraw@cs.umd.edu

Joel Tetreault
jtetreault@dataminr.com

Marine Carpuat
marine@cs.umd.edu

What is Style Transfer?

What is Style Transfer?

"style is an intuitive notion involving the manner in which something is said"

McDonald and Pustejovsky. 1985

What is Style Transfer?

“style is an intuitive notion involving the manner in which something is said”

McDonald and Pustejovsky. 1985

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

What is Style Transfer?

"style is an intuitive notion involving the manner in which something is said"

McDonald and Pustejovsky. 1985

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

What is Style Transfer?

"style is an intuitive notion involving the manner in which something is said"

McDonald and Pustejovsky. 1985

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

What is Style Transfer?

"style is an intuitive notion involving the manner in which something is said"

McDonald and Pustejovsky. 1985

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

What is Style Transfer?

"style is an intuitive notion involving the manner in which something is said"

McDonald and Pustejovsky. 1985

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

Informal

Gotta see both sides of the story

Formal

You have to consider both sides of the story.

Challenges in Style Transfer: Evaluation

Fluency **Meaning** **Style**

Generate a well-formed sentence that matches a desired stylistic attribute while preserving the meaning of the input sentence

Towards actual (not operational) textual style transfer auto-evaluation. Pang. **2019**
Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. Pang and Gimpel. **2019**
Evaluating style transfer for text. Mir et al. **2019**
Style transfer for texts: Retrain, report errors, compare with rewrites. Tikhonov et al. **2019**
Style transfer and paraphrase: Looking for a sensible semantic similarity metric. Yamshchikov et al. **2021**

Our work: Evaluating Automatic Metrics for Style Transfer Evaluation

(1) structured literature review

(2) empirical evaluation of most commonly used metrics

Our work: Evaluating Automatic Metrics for Style Transfer Evaluation

(1) structured literature review

(2) empirical evaluation of most commonly used metrics

Our work: Evaluating Automatic Metrics for Style Transfer Evaluation

(1) structured literature review

(2) empirical evaluation of most commonly used metrics

Our work: Evaluating Automatic Metrics for Style Transfer Evaluation

(1) structured literature review

(2) empirical evaluation of most commonly used metrics

Proposed best practices

- ✓ **Style:** XLM-R regression models fine-tuned on English
- ✓ **Meaning:** chRF score with input references
- ✓ **Fluency:** XLM-R pseudo-perplexity

Structured Review of ST evaluation:
Findings on Formality

PAPER ID	STYLE			MEANING			FLUENCY			OVERALL	
	metric	arch.		metric	arch.		metric	arch.		metric	
[1]	REG	Linear reg.	✗	CLS	CNN	✓	REG	Linear reg.	✗	r-BLEU	✓
[2]										r-BLEU	–
[3]	CLS	CNN	–	r-BLEU		–				GM(S,M)	–
[4]	CLS	CNN	✗	r-BLEU		✓				GM(S,M)	✓
[5]	CLS	CNN	✗	r-BLEU		✓					
[6]	CLS	LSTM	–	CLS	BERT	–				r-BLEU	–
[7]										r-BLEU	–
[8]	CLS	CNN	–								
[9]	CLS	LSTM	–	EMB-SIM		–	PPL	LM (RNN)	–	F1(S,M)	–
[10]	CLS	RoBERTa	✗	EMB-SIM		✓	PPL	LM (RoBERTa)	–	J(S,M,F)	✓
[11]	CLS	CNN	✗	r-BLEU		✗				F1(S,M)	–
[12]	CLS	GRU	✗				CLS	Linear reg.	✗	r-BLEU	✗
[13]	CLS	BERT	✓	r-BLEU		✓	PPL	LM (KenLM)	✗	GM(S,M,F)	–
[14]										r-BLEU	–
[15]	CLS	FASTTEXT	✓	r-BLEU		✓	PPL	LM (GPT)	✓		
[16]	CLS	CNN	–	r-BLEU		–	PPL	LM (LSTM)	–		
[17]	CLS	CNN	✓	r-BLEU		✓					
[18]	CLS	CNN	✓	r-BLEU		✓				GM|HM(S,M)	✓
[19]	CLS	GRU	–	CLS	BERT	–				r-BLEU	✓
[20]	CLS	RoBERTa	–	r-BLEU		–	PPL	LM (GPT)	–	GM|HM(S,M)	–
[21]	CLS	CNN	–	r-BLEU		✓	PPL	LM (GPT)	✓		
[22]										r-BLEU	–
[23]	REG	BERT	✗	s-BLEU		✓	PPL	LM (KenLM)	✗	r-BLEU	✗

Structured Review of ST evaluation:

Findings on Formality

PAPER ID	STYLE			MEANING			FLUENCY			OVERALL	
	metric	arch.		metric	arch.		metric	arch.		metric	
[1]	REG	Linear reg.	✗	CLS	CNN	✓	REG	Linear reg.	✗	r-BLEU	✓
[2]										r-BLEU	–
[3]	CLS	CNN	–	r-BLEU		–				GM(S,M)	–
[4]	CLS	CNN	✗	r-BLEU		✓				GM(S,M)	✓
[5]	CLS	CNN	✗	r-BLEU		✓					
[6]	CLS	LSTM	–	CLS	BERT	–				r-BLEU	–
[7]										r-BLEU	–
[8]	CLS	CNN	–								
[9]	CLS	LSTM	–	EMB-SIM		–	PPL	LM (RNN)	–	F1(S,M)	–
[10]	CLS	RoBERTa	✗	EMB-SIM		✓	PPL	LM (RoBERTa)	–	J(S,M,F)	✓
[11]	CLS	CNN	✗	r-BLEU		✗				F1(S,M)	–
[12]	CLS	GRU	✗				CLS	Linear reg.	✗	r-BLEU	✗
[13]	CLS	BERT	✓	r-BLEU		✓	PPL	LM (KenLM)	✗	GM(S,M,F)	–
[14]										r-BLEU	–
[15]	CLS	FASTTEXT	✓	r-BLEU		✓	PPL	LM (GPT)	✓		
[16]	CLS	CNN	–	r-BLEU		–	PPL	LM (LSTM)	–		
[17]	CLS	CNN	✓	r-BLEU		✓					
[18]	CLS	CNN	✓	r-BLEU		✓				GM|HM(S,M)	✓
[19]	CLS	GRU	–	CLS	BERT	–				r-BLEU	✓
[20]	CLS	RoBERTa	–	r-BLEU		–	PPL	LM (GPT)	–	GM|HM(S,M)	–
[21]	CLS	CNN	–	r-BLEU		✓	PPL	LM (GPT)	✓		
[22]										r-BLEU	–
[23]	REG	BERT	✗	s-BLEU		✓	PPL	LM (KenLM)	✗	r-BLEU	✗

Lack of **standardized** metrics

Structured Review of ST evaluation:
Findings on Formality

PAPER ID	STYLE			MEANING			FLUENCY			OVERALL	
	metric	arch.		metric	arch.		metric	arch.		metric	
[1]	REG	Linear reg.	✗	CLS	CNN	✓	REG	Linear reg.	✗	r-BLEU	✓
[2]										r-BLEU	–
[3]	CLS	CNN	–	r-BLEU		–				GM(S,M)	–
[4]	CLS	CNN	✗	r-BLEU		✓				GM(S,M)	✓
[5]	CLS	CNN	✗	r-BLEU		✓					
[6]	CLS	LSTM	–	CLS	BERT	–				r-BLEU	–
[7]										r-BLEU	–
[8]	CLS	CNN	–								
[9]	CLS	LSTM	–	EMB-SIM		–	PPL	LM (RNN)	–	F1(S,M)	–
[10]	CLS	RoBERTa	✗	EMB-SIM		✓	PPL	LM (RoBERTa)	–	J(S,M,F)	✓
[11]	CLS	CNN	✗	r-BLEU		✗				F1(S,M)	–
[12]	CLS	GRU	✗				CLS	Linear reg.	✗	r-BLEU	✗
[13]	CLS	BERT	✓	r-BLEU		✓	PPL	LM (KenLM)	✗	GM(S,M,F)	–
[14]										r-BLEU	–
[15]	CLS	FASTTEXT	✓	r-BLEU		✓	PPL	LM (GPT)	✓		
[16]	CLS	CNN	–	r-BLEU		–	PPL	LM (LSTM)	–		
[17]	CLS	CNN	✓	r-BLEU		✓					
[18]	CLS	CNN	✓	r-BLEU		✓				GM∣HM(S,M)	✓
[19]	CLS	GRU	–	CLS	BERT	–				r-BLEU	✓
[20]	CLS	RoBERTa	–	r-BLEU		–	PPL	LM (GPT)	–	GM∣HM(S,M)	–
[21]	CLS	CNN	–	r-BLEU		✓	PPL	LM (GPT)	✓		
[22]										r-BLEU	–
[23]	REG	BERT	✗	s-BLEU		✓	PPL	LM (KenLM)	✗	r-BLEU	✗

- Lack of **standardized** metrics
- Lack of **agreement with human** judgments

Structured Review of ST evaluation:
Findings on Formality

PAPER ID	STYLE			MEANING			FLUENCY			OVERALL	
	metric	arch.		metric	arch.		metric	arch.		metric	
[1]	REG	Linear reg.	✗	CLS	CNN	✓	REG	Linear reg.	✗	r-BLEU	✓
[2]										r-BLEU	–
[3]	CLS	CNN	–	r-BLEU		–				GM(S,M)	–
[4]	CLS	CNN	✗	r-BLEU		✓				GM(S,M)	✓
[5]	CLS	CNN	✗	r-BLEU		✓					
[6]	CLS	LSTM	–	CLS	BERT	–				r-BLEU	–
[7]										r-BLEU	–
[8]	CLS	CNN	–								
[9]	CLS	LSTM	–	EMB-SIM		–	PPL	LM (RNN)	–	F1(S,M)	–
[10]	CLS	RoBERTa	✗	EMB-SIM		✓	PPL	LM (RoBERTa)	–	J(S,M,F)	✓
[11]	CLS	CNN	✗	r-BLEU		✗				F1(S,M)	–
[12]	CLS	GRU	✗				CLS	Linear reg.	✗	r-BLEU	✗
[13]	CLS	BERT	✓	r-BLEU		✓	PPL	LM (KenLM)	✗	GM(S,M,F)	–
[14]										r-BLEU	–
[15]	CLS	FASTTEXT	✓	r-BLEU		✓	PPL	LM (GPT)	✓		
[16]	CLS	CNN	–	r-BLEU		–	PPL	LM (LSTM)	–		
[17]	CLS	CNN	✓	r-BLEU		✓					
[18]	CLS	CNN	✓	r-BLEU		✓				GM|HM(S,M)	✓
[19]	CLS	GRU	–	CLS	BERT	–				r-BLEU	✓
[20]	CLS	RoBERTa	–	r-BLEU		–	PPL	LM (GPT)	–	GM|HM(S,M)	–
[21]	CLS	CNN	–	r-BLEU		✓	PPL	LM (GPT)	✓		
[22]										r-BLEU	–
[23]	REG	BERT	✗	s-BLEU		✓	PPL	LM (KenLM)	✗	r-BLEU	✗

- Lack of **standardized** metrics
- Lack of **agreement with human** judgments
- Lack of **portability to languages** other than English

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation

Fluency

Style

Meaning

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation

Fluency

Style

Meaning

Correlation analysis of automatic metrics w/ human ratings

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation

Fluency | Style | Meaning

Correlation analysis of automatic metrics w/ human ratings

… through a multilingual lens

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation

Fluency | Style | Meaning

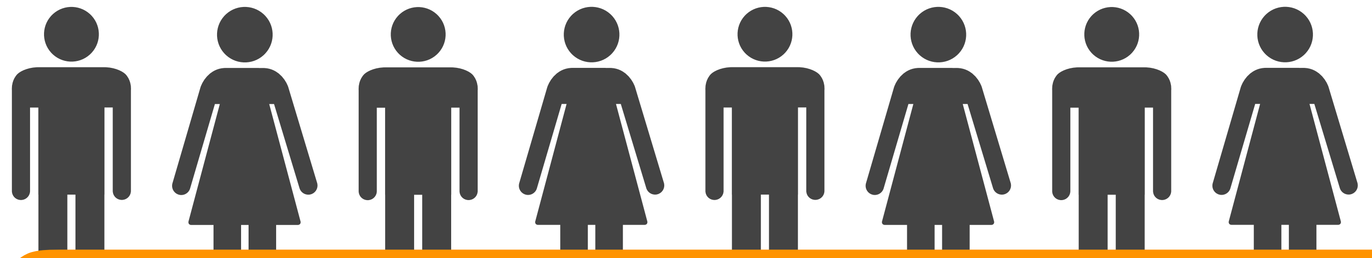
- Correlation analysis of automatic metrics w/ human ratings
- … through a multilingual lens
- … with formality as a case study

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation: Formality focus

Availability of human ratings collected consistently across evaluation dimensions multiple languages

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation: Formality focus

Availability of human ratings collected consistently across evaluation dimensions multiple languages*



Rate the fluency of the given sentence from 1 to 5

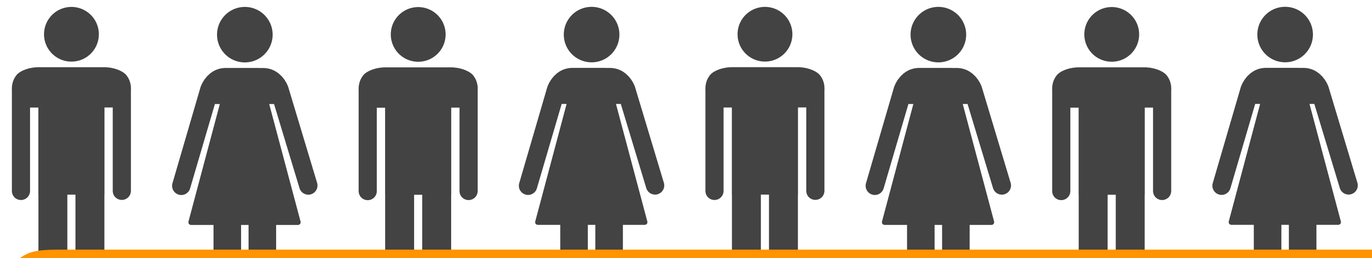
Rate the formality of the given sentence from -3 to +3

Rate the similarity of the two sentences from 1 to 6

*Rao et al, Briakou et al.

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation: Formality focus

Availability of human ratings collected consistently across evaluation dimensions multiple languages *



Rate the fluency of the given sentence from 1 to 5

Rate the formality of the given sentence from -3 to +3

Rate the similarity of the two sentences from 1 to 6

- English
- Brazilian-Portuguese
- Italian
- French

* Rao et al, Briakou et al.

Empirical Evaluation of Automatic Metrics for Style Transfer Evaluation: Formality focus

Availability of human ratings collected consistently across evaluation dimensions multiple languages*

- Rate the fluency of the given sentence from 1 to 5
- Rate the formality of the given sentence from -3 to +3
- Rate the similarity of the two sentences from 1 to 6

English

Brazilian-Portuguese

Italian

French

5 systems per language; 100-500 outputs per system

*Rao et al, Briakou et al.

Automatic Evaluation for Formality:
Prior work uses lots of different approaches

Linear regressor

LSTM classifier

FASTTEXT classifier

GRU classifier

RoBerta classifier

CNN classifier

BERT classifier

BERT regressor

Automatic Evaluation for Formality: Prior work uses lots of different approaches

Linear regressor

LSTM classifier

FASTTEXT classifier

GRU classifier

RoBerta classifier

CNN classifier

BERT classifier

BERT regressor

Automatic Evaluation for Formality: Dimensions of comparison

Linear regressor
LSTM classifier
FASTTEXT classifier
GRU classifier
RoBerta classifier
CNN classifier
BERT classifier
BERT regressor

(1) Task framing: regression vs. classification

Automatic Evaluation for Formality: Dimensions of comparison

(1) Task framing: regression vs. classification

(2) Multilingual framing: cross-lingual transfer

Automatic Evaluation for Formality: Experimental Settings

➡ **Models:**
Fine-tuning multilingual pre-trained models (i.e., mBERT vs. XLM-R)

➡ **Cross-lingual Transfer:**

- TRANSLATE-TRAIN
- TRANSLATE-TEST
- ZERO-SHOT

Automatic Evaluation for Formality: Experimental Settings

- **Models:**
  Fine-tuning multilingual pre-trained models (i.e., mBERT vs. XLM-R)

- **Cross-lingual Transfer:**
  - TRANSLATE-TRAIN
  - TRANSLATE-TEST
  - ZERO-SHOT

TRAINING DATA

EN Training data

Machine Translate in i.e., FR

INFERENCE DATA

Fr Test data

Automatic Evaluation for Formality: Experimental Settings

- **Models:**
  Fine-tuning multilingual pre-trained models (i.e., mBERT vs. XLM-R)
- **Cross-lingual Transfer:**
  - TRANSLATE-TRAIN
  - TRANSLATE-TEST
  - ZERO-SHOT

TRAINING DATA

EN Training data

INFERENCE DATA

Fr Test data

Machine Translate in EN

Automatic Evaluation for Formality: Experimental Settings

- **Models:**
  Fine-tuning multilingual pre-trained models (i.e., mBERT vs. XLM-R)
- **Cross-lingual Transfer:**
  - TRANSLATE-TRAIN
  - TRANSLATE-TEST
  - ZERO-SHOT

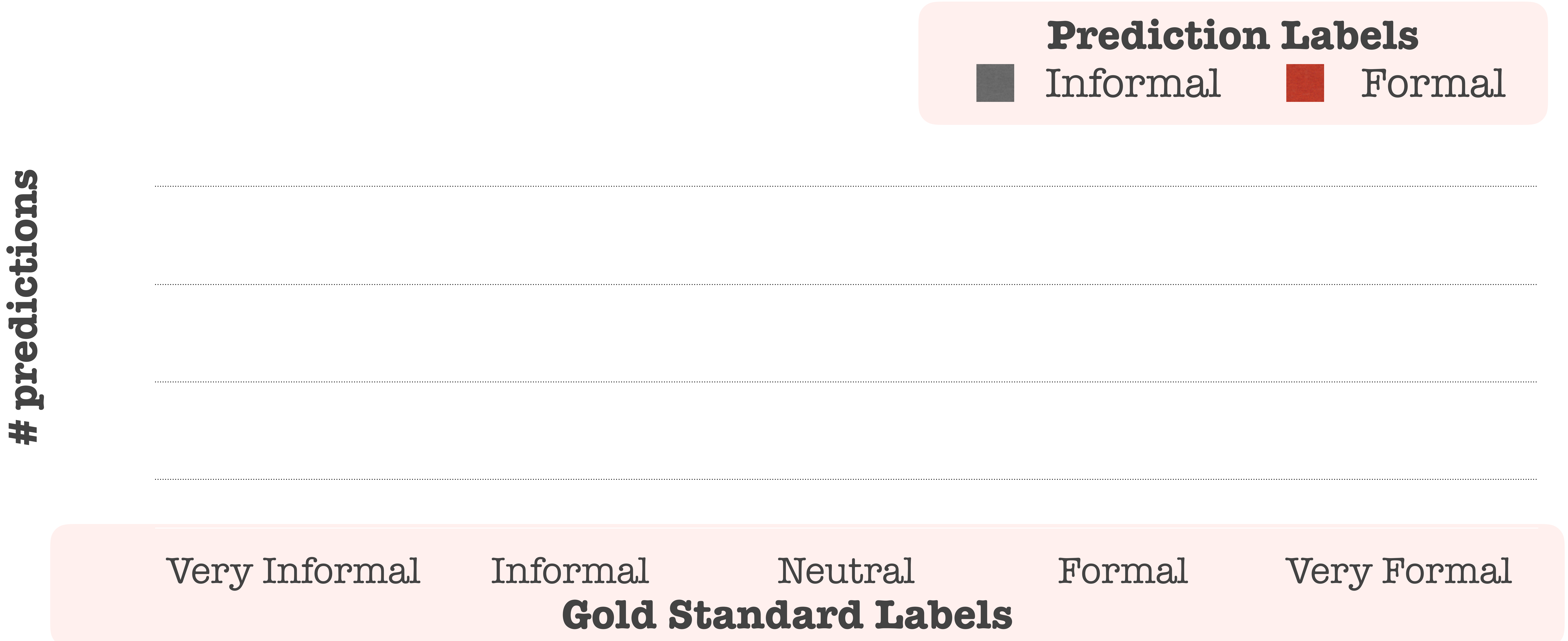
TRAINING DATA

EN Training data

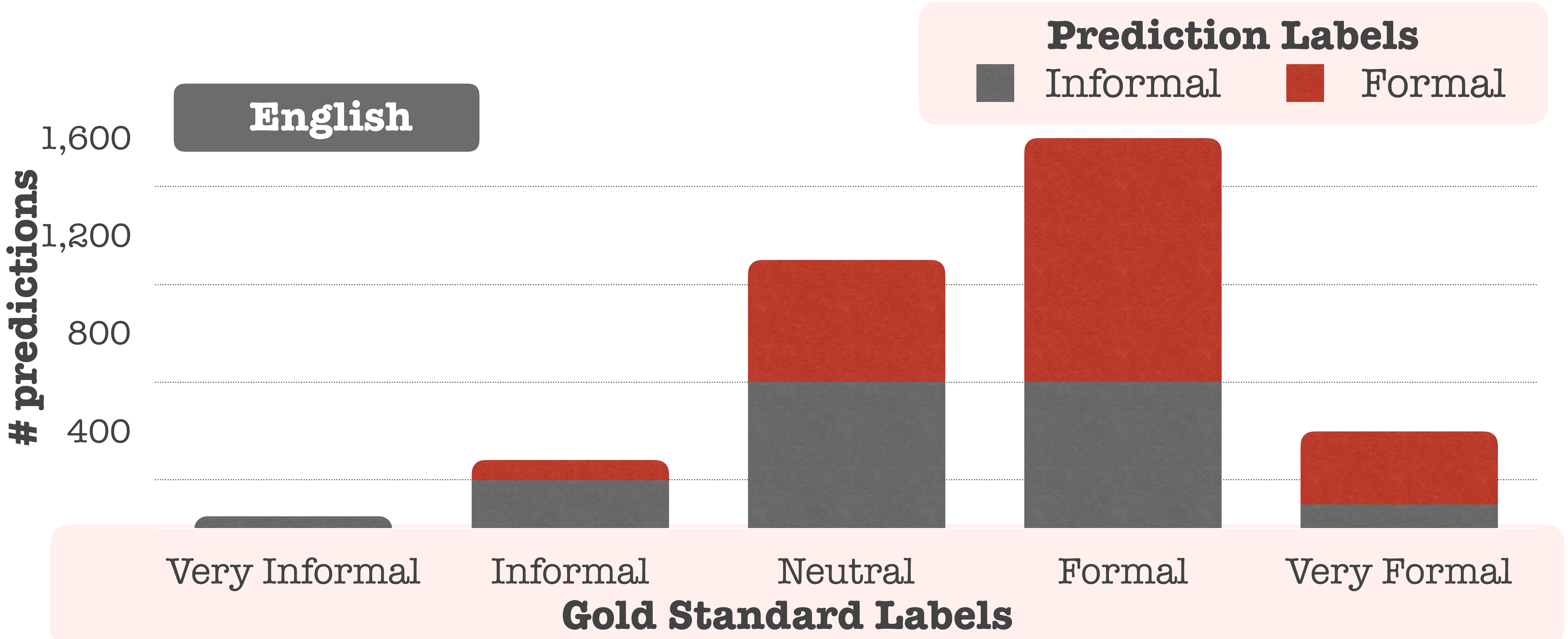
INFERENCE DATA

Fr Test data

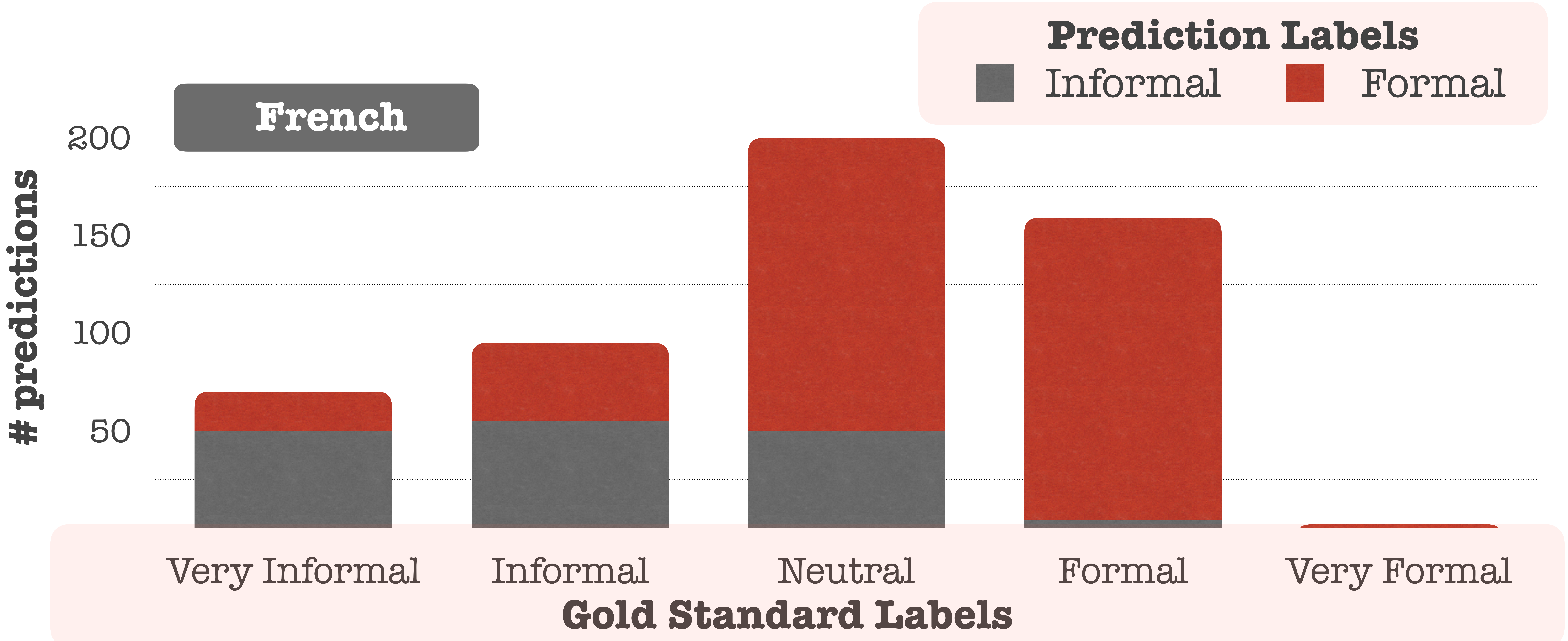
Evaluating Binary Formality Classifiers



Binary Formality Classifiers...
Lack sensitivity to different formality levels



Binary Formality Classifiers...
Are biased towards the formal class



Evaluating Formality Regressors

ZERO-SHOT

Best Practice for Formality Evaluation:

XLM-R regressor in ZERO-SHOT setting

Best Practice for Formality Evaluation:

XLM-R regressor in ZERO-SHOT setting

Best Practice for Formality Evaluation:

XLM-R regressor in ZERO-SHOT setting

Automatic Evaluation for Meaning: Experimental Settings

String-based
Require access to a reference segment

s-BLEU
chrF
METEOR
r-BLEU

Supervised
Fine-tune on labeled data
(Semantic Textual Similarity)

mBERT
XML-R
X-transfer

Unsupervised
Based on pre-trained embeddings

BERT-score
Word's Movers Distance

Automatic Evaluation for Meaning: Experimental Settings

String-based
Require access to a reference segment

s-BLEU chrF
METEOR r-BLEU

Supervised
Fine-tune on labeled data
(Semantic Textual Similarity)

mBERT XML-R
X-transfer

Unsupervised
Based on pre-trained embeddings

BERT-score
Word's Movers Distance

Automatic Evaluation for Meaning: Experimental Settings

String-based
Require access to a reference segment

s-BLEU chrF
METEOR r-BLEU

Supervised
Fine-tune on labeled data
(Semantic Textual Similarity)

mBERT XML-R
X-transfer

Unsupervised
Based on pre-trained embeddings

BERT-score
Word's Movers Distance

Automatic Evaluation for Meaning: Experimental Settings

String-based
Require access to a reference segment

s-BLEU chrF
METEOR r-BLEU

Supervised
Fine-tune on labeled data
(Semantic Textual Similarity)

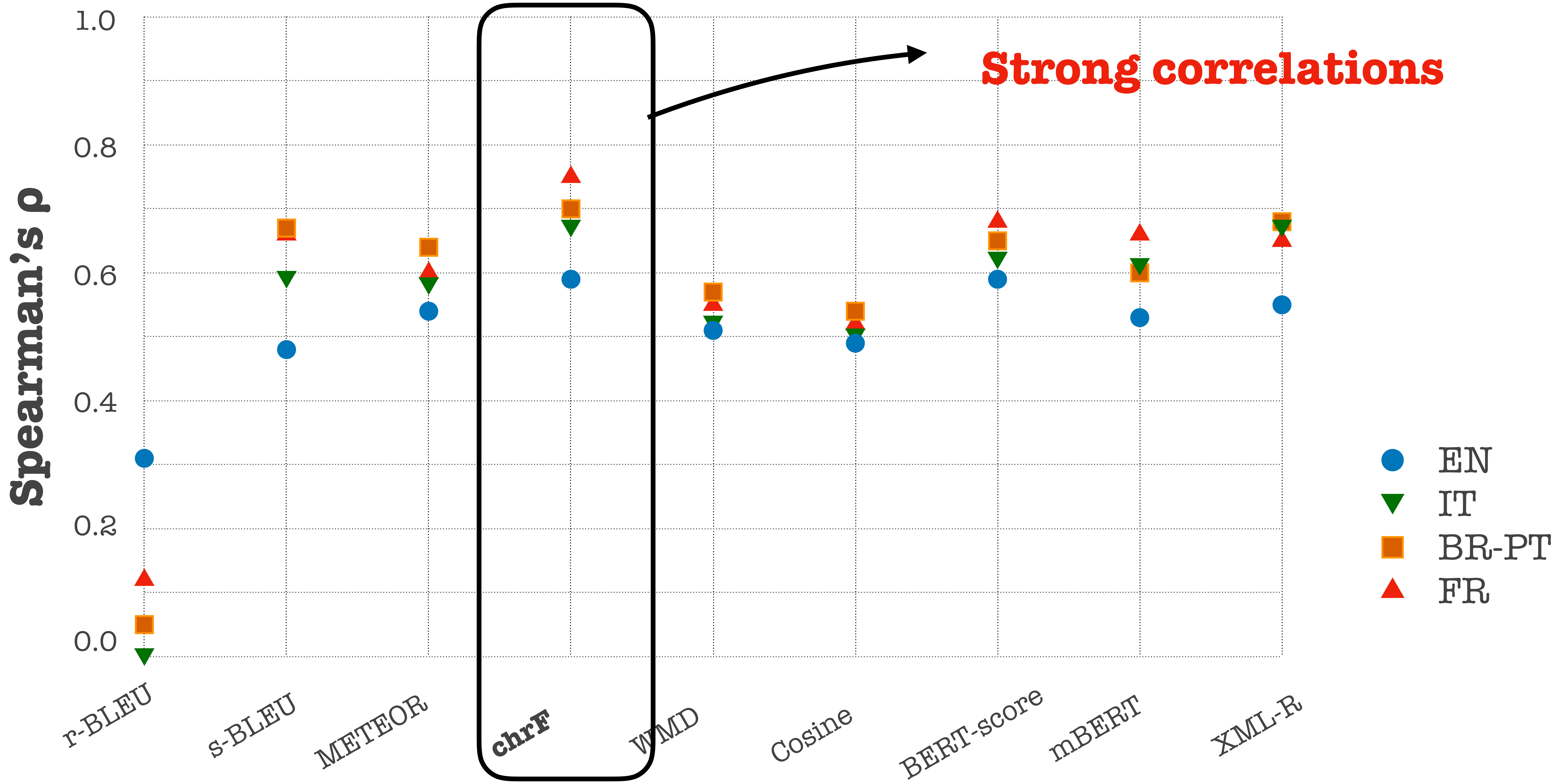
mBERT XML-R
X-transfer

Unsupervised
Based on pre-trained embeddings

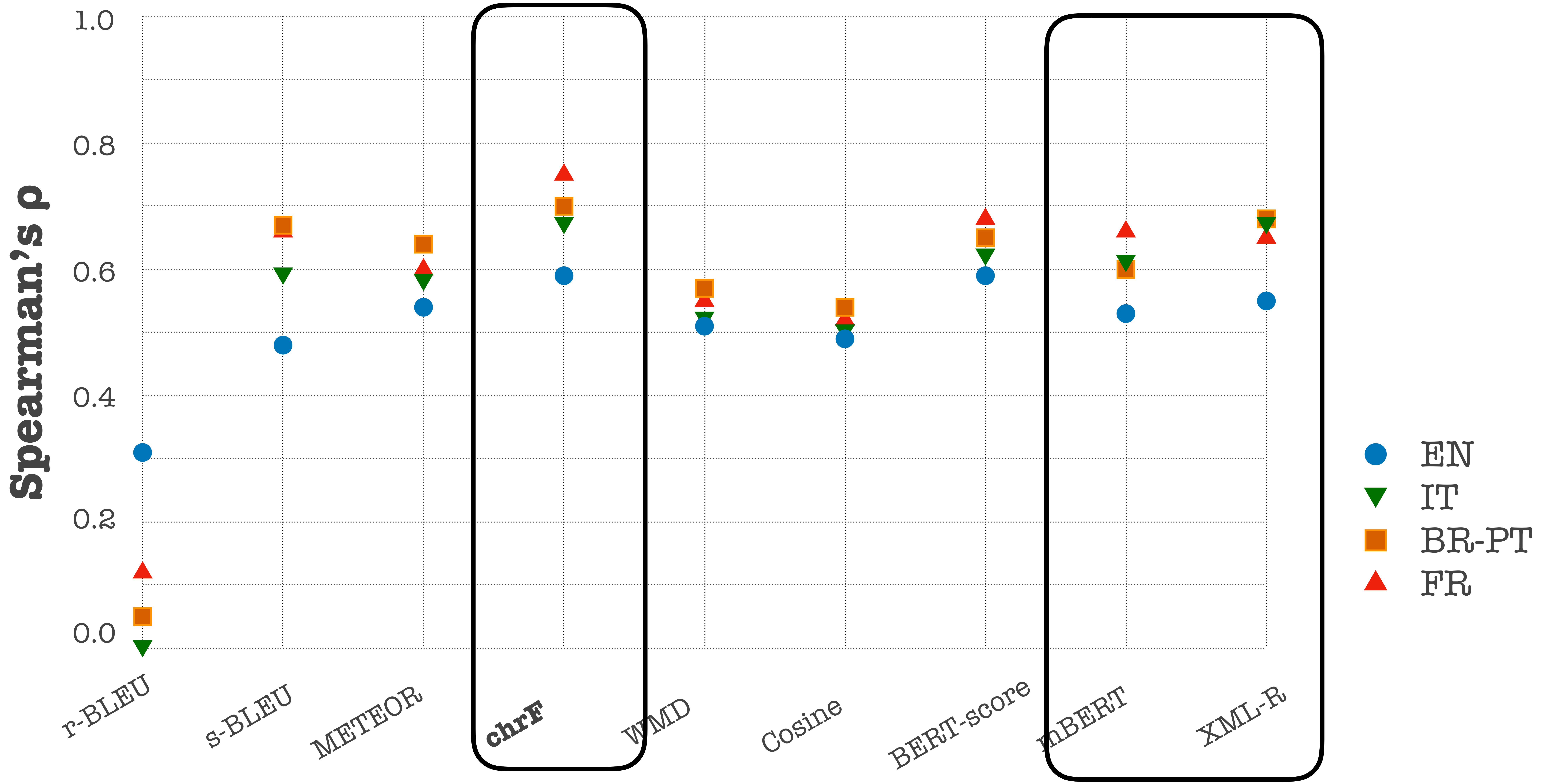
Cosine BERT-score
Word's Movers Distance

Best Practice for Meaning Evaluation:

chrF yields strong correlations across languages

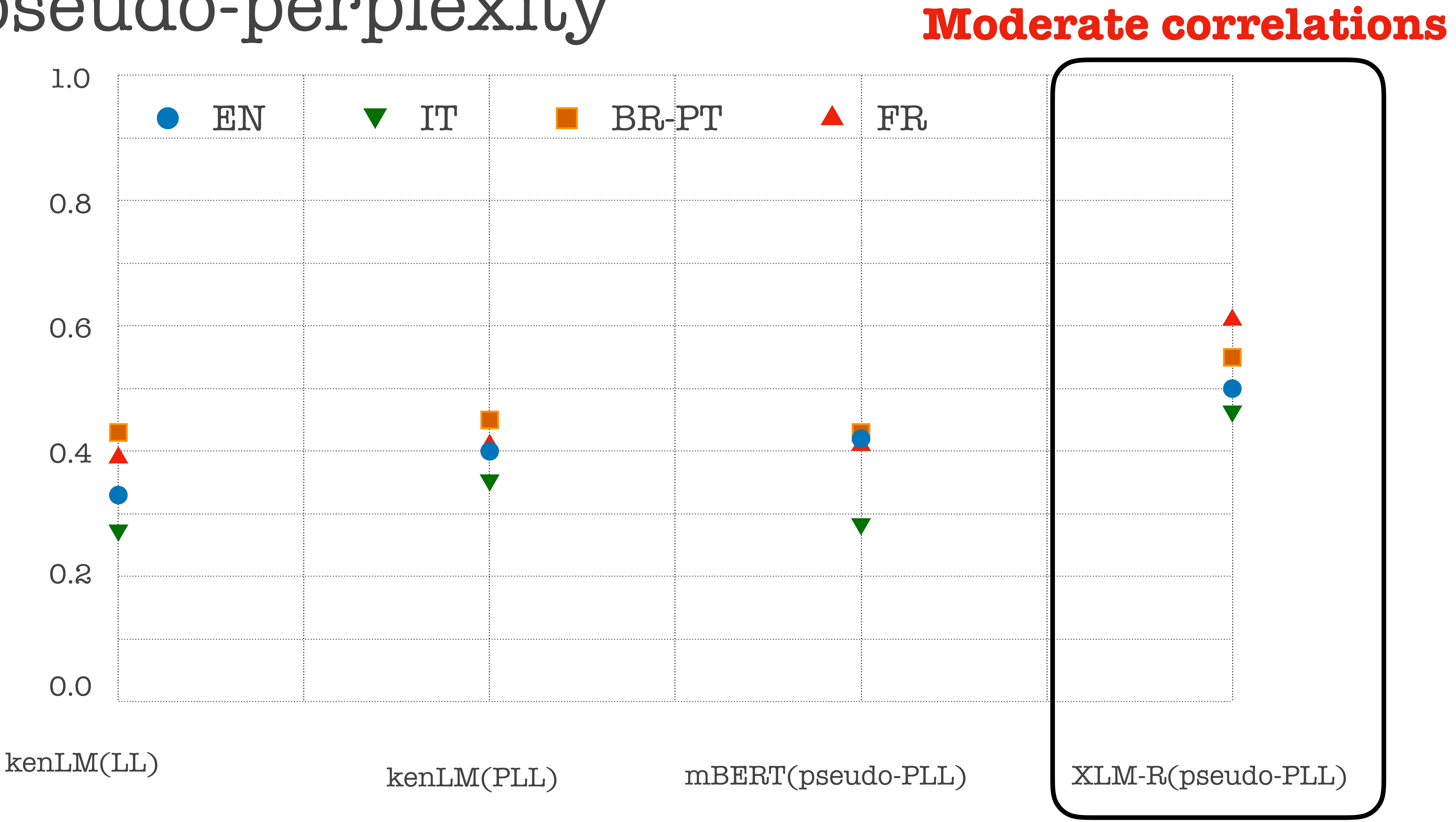


Best Practice for Meaning Evaluation:
chrF yields strong correlations across languages



Best Practice for Fluency Evaluation:

XLM-R pseudo-perplexity



Summary of Findings

Limitations of current automatic evaluation for FoST

- Lack of standardized metrics
- Lack of agreement with human judgments
- Lack of portability to languages other than English

Proposed best practices

- ✓ **Style:** XLM-R regression models fine-tuned on English
- ✓ **Meaning:** chRF score with input references
- ✓ **Fluency:** XLM-R pseudo-perplexity

Code & Data: https://github.com/ Elbria/xformal-FoST-meta

Summary of Findings

Limitations of current automatic evaluation for FoST

- Lack of standardized metrics
- Lack of agreement with human judgments
- Lack of portability to languages other than English

Proposed best practices

- ✓ **Style:** XLM-R regression models fine-tuned on English
- ✓ **Meaning:** chRF score with input references
- ✓ **Fluency:** XLM-R pseudo-perplexity

Code & Data: https://github.com/ Elbria/xformal-FoST-meta

Summary of Findings

Limitations of current automatic evaluation for FoST

- Lack of standardized metrics
- Lack of agreement with human judgments
- Lack of portability to languages other than English

Proposed best practices

✓ **Style:** XLM-R regression models fine-tuned on English
✓ **Meaning:** chRF score with input references
✓ **Fluency:** XLM-R pseudo-perplexity

Code & Data: https://github.com/ Elbria/xformal-FoST-meta