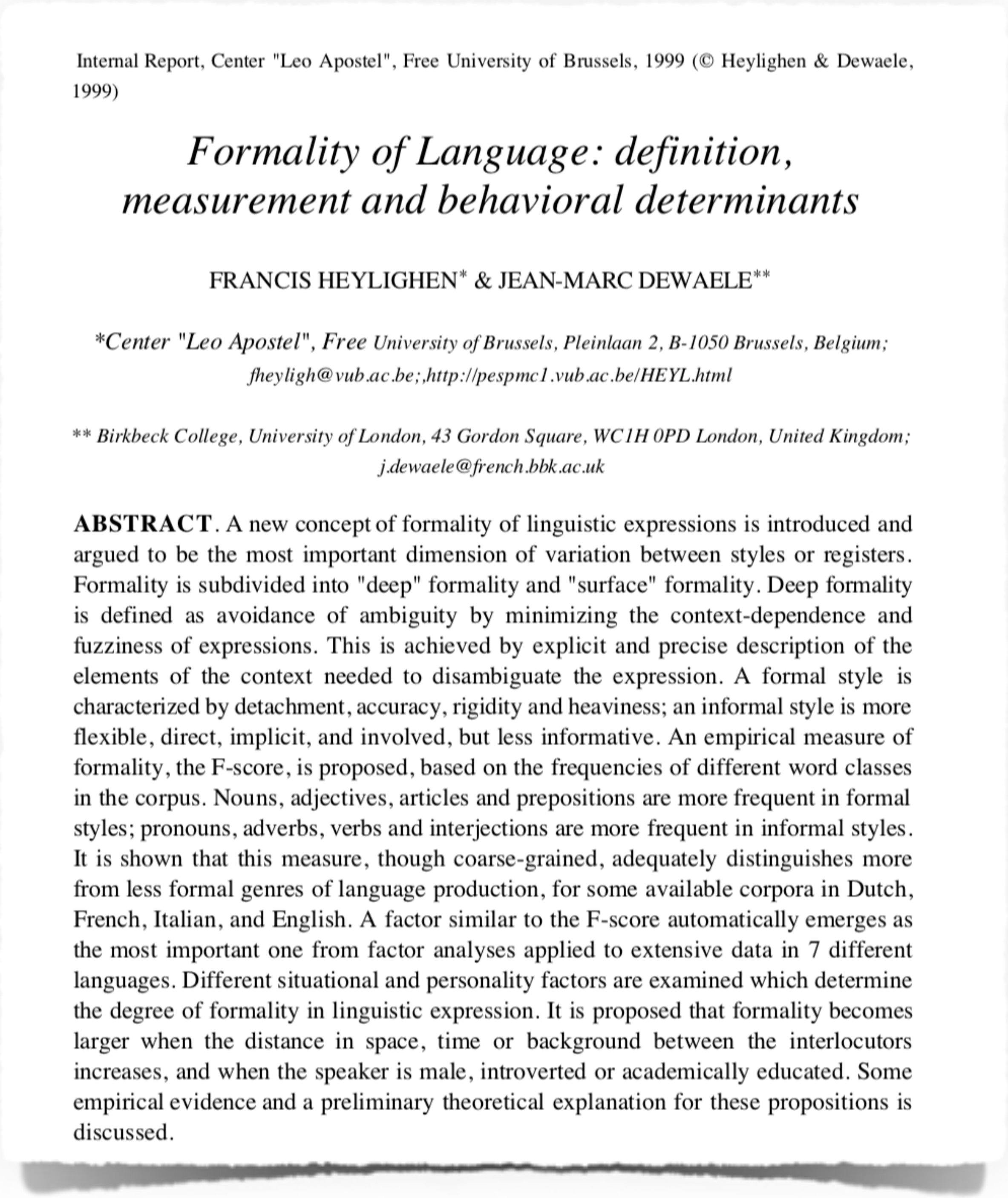
Tracking Progress in Style Transfer: from Human to Automatic Evaluation

Eleftheria Briakou
ebriakou@umd.edu



Internal Report, Center "Leo Apostel", Free University of Brussels, 1999 (© Heylighen & Dewaele, 1999)

Formality of Language: definition, measurement and behavioral determinants

FRANCIS HEYLIGHEN* & JEAN-MARC DEWAELE**

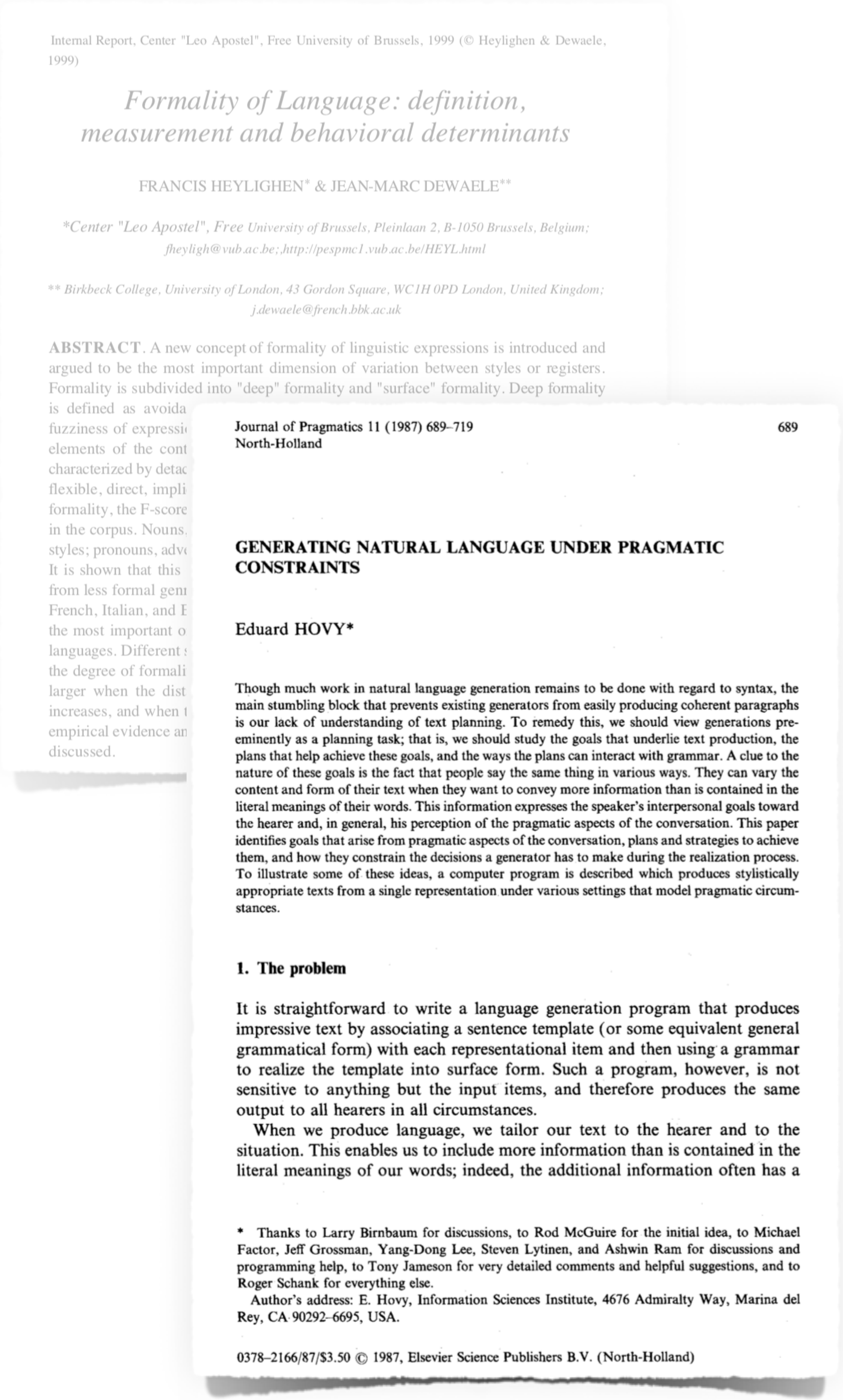
*Center "Leo Apostel", Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium; fheyligh@vub.ac.be; http://pespmc1.vub.ac.be/HEYL.html

** Birkbeck College, University of London, 43 Gordon Square, WC1H 0PD London, United Kingdom; j.dewaele@french.bbk.ac.uk

ABSTRACT. A new concept of formality of linguistic expressions is introduced and argued to be the most important dimension of variation between styles or registers. Formality is subdivided into "deep" formality and "surface" formality. Deep formality is defined as avoidance of ambiguity by minimizing the context-dependence and fuzziness of expressions. This is achieved by explicit and precise description of the elements of the context needed to disambiguate the expression. A formal style is characterized by detachment, accuracy, rigidity and heaviness; an informal style is more flexible, direct, implicit, and involved, but less informative. An empirical measure of formality, the F-score, is proposed, based on the frequencies of different word classes in the corpus. Nouns, adjectives, articles and prepositions are more frequent in formal styles; pronouns, adverbs, verbs and interjections are more frequent in informal styles. It is shown that this measure, though coarse-grained, adequately distinguishes more from less formal genres of language production, for some available corpora in Dutch, French, Italian, and English. A factor similar to the F-score automatically emerges as the most important one from factor analyses applied to extensive data in 7 different languages. Different situational and personality factors are examined which determine the degree of formality in linguistic expression. It is proposed that formality becomes larger when the distance in space, time or background between the interlocutors increases, and when the speaker is male, introverted or academically educated. Some empirical evidence and a preliminary theoretical explanation for these propositions is discussed.

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

McDonald and Pustejovsky. 1985


Internal Report, Center "Leo Apostel", Free University of Brussels, 1999 (© Heylighen & Dewaele, 1999)

Formality of Language: definition, measurement and behavioral determinants

FRANCIS HEYLIGHEN* & JEAN-MARC DEWAELE**

*Center "Leo Apostel", Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium; fheyligh@vub.ac.be; http://pespmc1.vub.ac.be/HEYL.html

** Birkbeck College, University of London, 43 Gordon Square, WC1H 0PD London, United Kingdom; j.dewaele@french.bbk.ac.uk

ABSTRACT. A new concept of formality of linguistic expressions is introduced and argued to be the most important dimension of variation between styles or registers. Formality is subdivided into "deep" formality and "surface" formality. Deep formality
is defined as avoida
fuzziness of expressi
elements of the cont
characterized by detac
flexible, direct, impli
formality, the F-score
in the corpus. Nouns
styles; pronouns, adv
It is shown that this
from less formal gen
French, Italian, and E
the most important o
languages. Different
the degree of formali
larger when the dist
increases, and when t
empirical evidence an
discussed.

Journal of Pragmatics 11 (1987) 689–719
North-Holland

689

GENERATING NATURAL LANGUAGE UNDER PRAGMATIC CONSTRAINTS

Eduard HOVY*

Though much work in natural language generation remains to be done with regard to syntax, the main stumbling block that prevents existing generators from easily producing coherent paragraphs is our lack of understanding of text planning. To remedy this, we should view generations pre-eminently as a planning task; that is, we should study the goals that underlie text production, the plans that help achieve these goals, and the ways the plans can interact with grammar. A clue to the nature of these goals is the fact that people say the same thing in various ways. They can vary the content and form of their text when they want to convey more information than is contained in the literal meanings of their words. This information expresses the speaker's interpersonal goals toward the hearer and, in general, his perception of the pragmatic aspects of the conversation. This paper identifies goals that arise from pragmatic aspects of the conversation, plans and strategies to achieve them, and how they constrain the decisions a generator has to make during the realization process. To illustrate some of these ideas, a computer program is described which produces stylistically appropriate texts from a single representation under various settings that model pragmatic circumstances.

1. The problem

It is straightforward to write a language generation program that produces impressive text by associating a sentence template (or some equivalent general grammatical form) with each representational item and then using a grammar to realize the template into surface form. Such a program, however, is not sensitive to anything but the input items, and therefore produces the same output to all hearers in all circumstances.

When we produce language, we tailor our text to the hearer and to the situation. This enables us to include more information than is contained in the literal meanings of our words; indeed, the additional information often has a

* Thanks to Larry Birnbaum for discussions, to Rod McGuire for the initial idea, to Michael Factor, Jeff Grossman, Yang-Dong Lee, Steven Lytinen, and Ashwin Ram for discussions and programming help, to Tony Jameson for very detailed comments and helpful suggestions, and to Roger Schank for everything else.
Author's address: E. Hovy, Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292–6695, USA.

0378–2166/87/$3.50 © 1987, Elsevier Science Publishers B.V. (North-Holland)

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

McDonald and Pustejovsky. 1985

"when we produce language,
we tailor our text to the hearer/situation"

Hovy. 1987

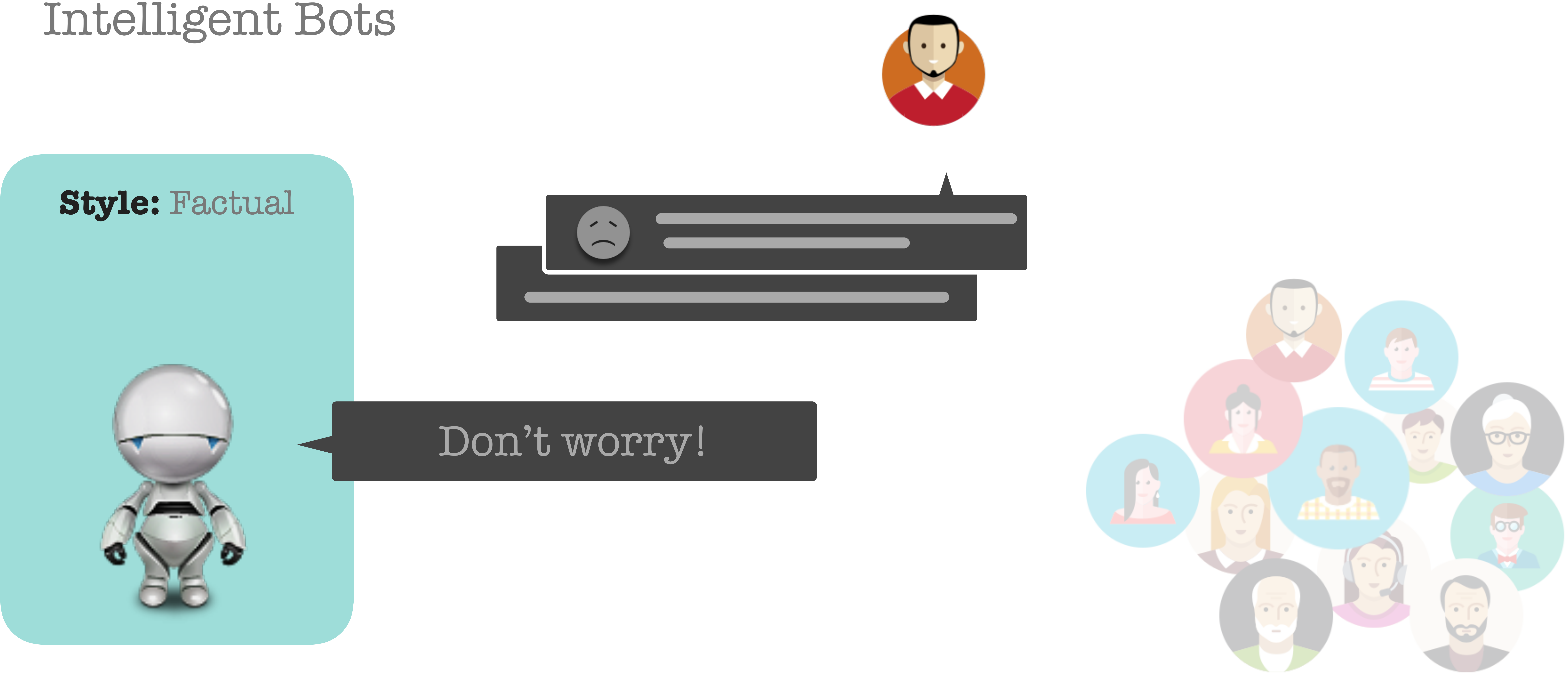
Transforming Style with Artificial Intelligence

Transforming Style with Artificial Intelligence

Intelligent Bots

Transforming Style with Artificial Intelligence

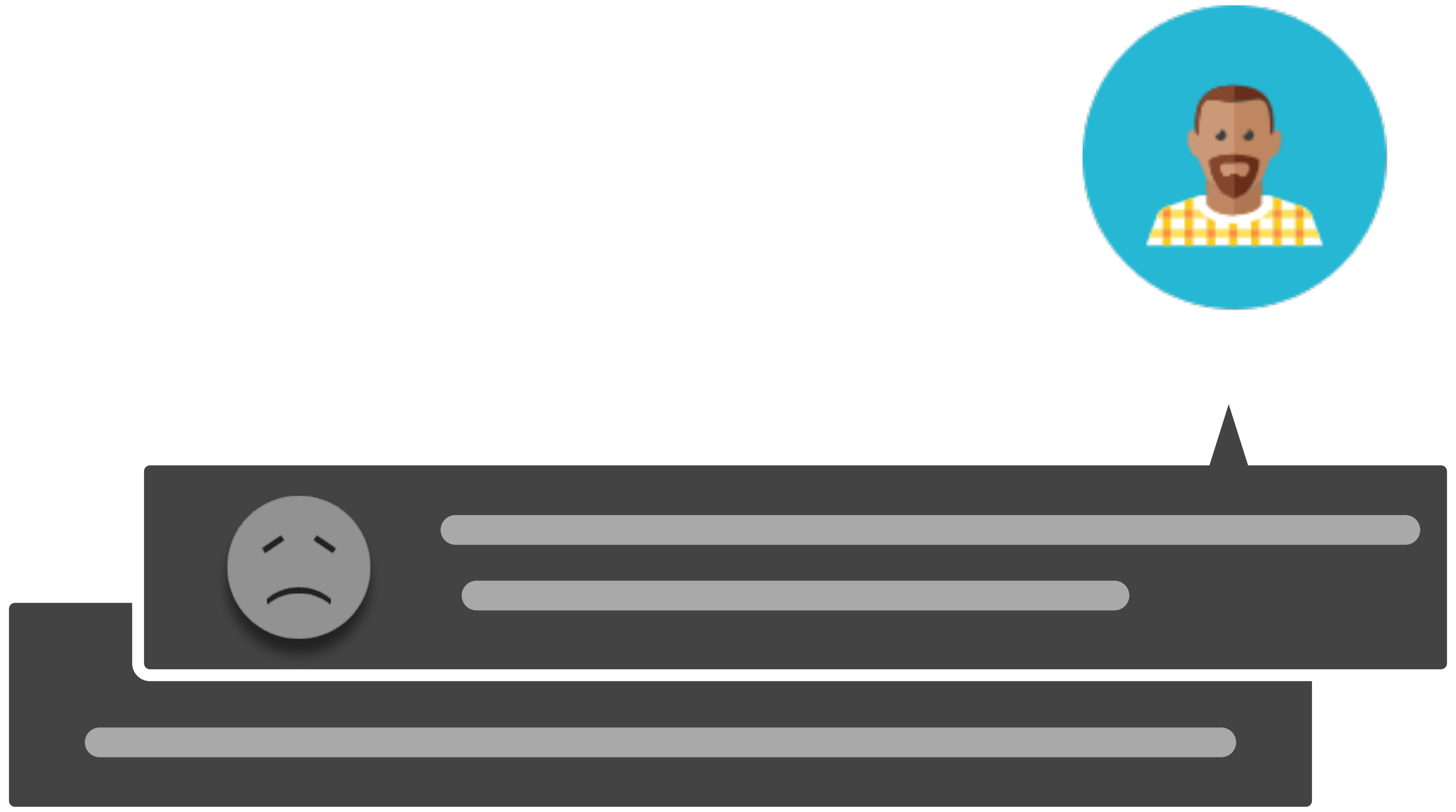
Intelligent Bots



Transforming Style with Artificial Intelligence

Intelligent Bots

Style: Empathetic



I’m sorry to hear this is troubling you!

Transforming Style with Artificial Intelligence

Intelligent Bots

Intelligent Writing/Teaching Assistants

Gotta see both sides of the story

FORMALITY

~~Gotta~~ → Have to Must

The use of slang such as ***Gotta*** may not be appropriate in this context. Consider using a standard word or phrase instead.

Transforming Style with Artificial Intelligence

Intelligent Bots

Intelligent Writing/Teaching Assistants

Mitigating Social Issues

Transforming Style with Artificial Intelligence

Intelligent Bots

Intelligent Writing/Teaching Assistants

Mitigating Social Issues

- Fighting offensive languages

Transforming Style with Artificial Intelligence

Intelligent Bots

Intelligent Writing/Teaching Assistants

Mitigating Social Issues

- Fighting offensive languages
- De-biasing online text

Transforming Style with Artificial Intelligence

Intelligent Bots

Intelligent Writing/Teaching Assistants

Mitigating Social Issues

- Fighting offensive languages
- De-biasing online text

...

Textual Style Transfer

Textual Style Transfer

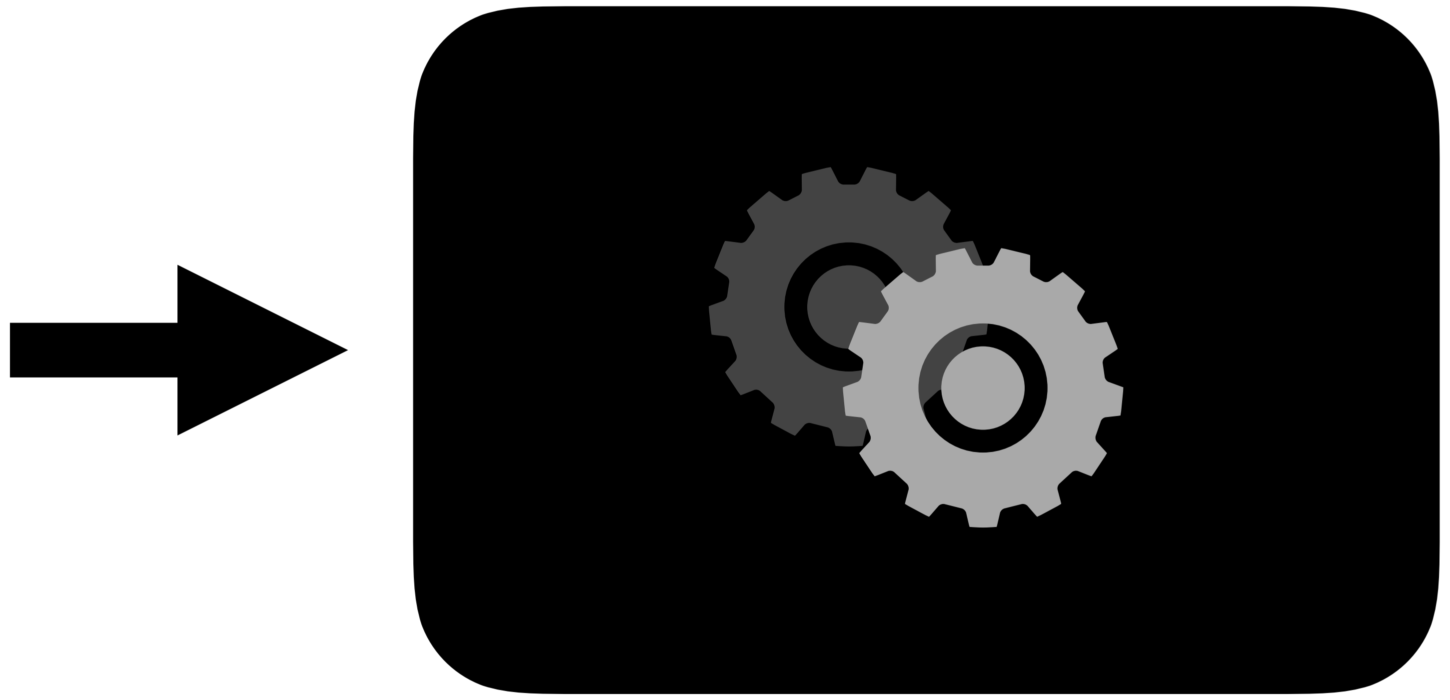
Inputs

Sentence	Gotta see both sides of the story
Target Style	Formal

Textual Style Transfer

Inputs

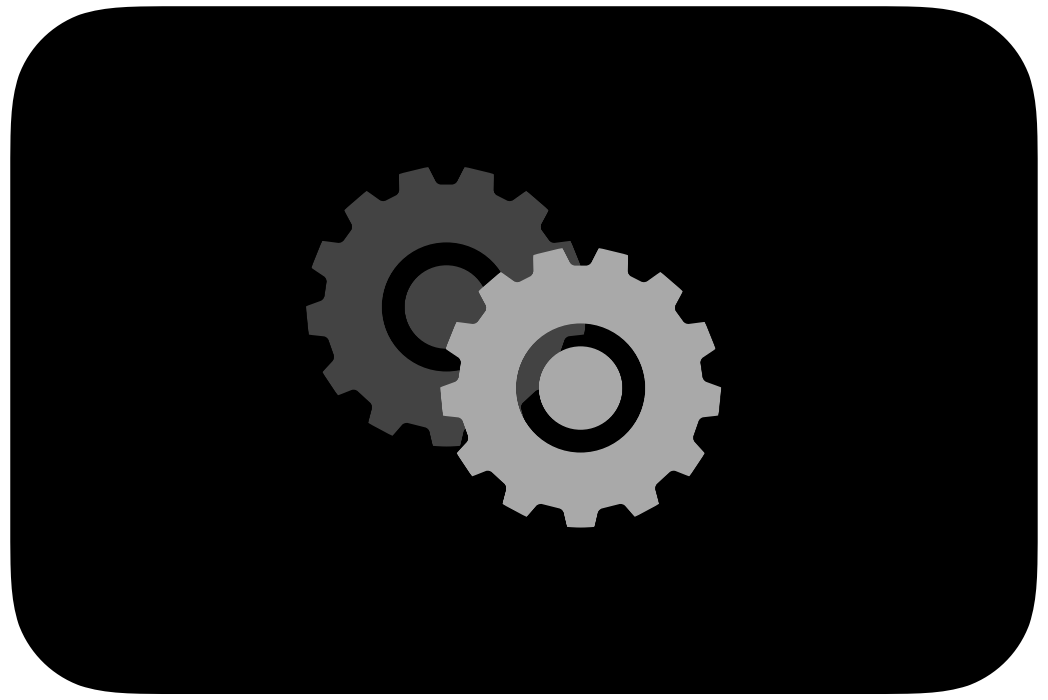
Sentence	Gotta see both sides of the story
Target Style	Formal



Textual Style Transfer

Inputs

Sentence	Gotta see both sides of the story
Target Style	Formal



Output

Sentence	You have to consider both sides of the story.

EVALUATION Dimensions in Style Transfer

EVALUATION Dimensions in Style Transfer

Properties of output:

Evaluation Dimensions:

EVALUATION Dimensions in Style Transfer

Properties of output:

well-formed sentence

matches a desired stylistic attribute

preserving the meaning of input

Evaluation Dimensions:

Fluency

EVALUATION Dimensions in Style Transfer

Properties of output:

well-formed sentence

matches a desired stylistic attribute

preserving the meaning of input

Evaluation Dimensions:

- Fluency
- Style

EVALUATION Dimensions in Style Transfer

Properties of output:

well-formed sentence

matches a desired stylistic attribute

preserving the meaning of input

Evaluation Dimensions:

- **Fluency** ☑
- **Style** ☑
- **Meaning** ☑

Challenges in Style Transfer EVALUATION

Alexey Tikhonov, Viacheslav Shibaev. Aleksander Nagaev, Aigul Nugmanova & Ivan P. Yamshchikov
Style transfer for texts: Retrain, report errors, compare with rewrites.
In proceedings of EMNLP 2019

Richard Yuanzhe Pang & Kevin Gimpel
Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer.
In Proceedings of 3rd NGT Workshop (@EMNLP) 2019

Remi Mir, Bjarke Felbo, Nick Obradovich & Iyad Rahwan
Evaluating style transfer for text.
In Proceedings of NAACL 2019

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov & Alexey Tikhonov
Style transfer and paraphrase: Looking for a sensible semantic similarity metric.
In Proceedings of AAAI 2021

Challenges in Style Transfer EVALUATION

Standard metrics for style accuracy
& meaning preservation **vary** across reruns!

Alexey Tikhonov, Viacheslav Shibaev. Aleksander Nagaev, Aigul Nugmanova & Ivan P. Yamshchikov

Style transfer for texts: Retrain, report errors, compare with rewrites.

In proceedings of EMNLP 2019

Challenges in Style Transfer EVALUATION

Standard metrics for style accuracy
& meaning preservation **vary** across reruns!

Alexey Tikhonov, Viacheslav Shibaev. Aleksander Nagaev, Aigul Nugmanova & Ivan P. Yamshchikov

Style transfer for texts: Retrain, report errors, compare with rewrites.

In proceedings of EMNLP 2019

Summarizing multiple metrics in one score
remains an open problem.

Richard Yuanzhe Pang & Kevin Gimpel

Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer.

In Proceedings of 3rd NGT Workshop (@EMNLP) 2019

Challenges in Style Transfer **EVALUATION**

Standard metrics for style accuracy
& meaning preservation **vary** across reruns!

Alexey Tikhonov, Viacheslav Shibaev. Aleksander Nagaev, Aigul Nugmanova & Ivan P. Yamshchikov

Style transfer for texts: Retrain, report errors, compare with rewrites.

In proceedings of EMNLP 2019

Summarizing multiple metrics in one score remains an open problem.

Richard Yuanzhe Pang & Kevin Gimpel

Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer.

In Proceedings of 3rd NGT Workshop (@EMNLP) 2019

Remi Mir, Bjarke Felbo, Nick Obradovich & Iyad Rahwan

Evaluating style transfer for text.

In Proceedings of NAACL 2019

Models exhibit trafeoffs between evaluation dimensions

Challenges in Style Transfer EVALUATION

Standard metrics for style accuracy
& meaning preservation **vary** across reruns!

Alexey Tikhonov, Viacheslav Shibaev. Aleksander Nagaev, Aigul Nugmanova & Ivan P. Yamshchikov

Style transfer for texts: Retrain, report errors, compare with rewrites.

In proceedings of EMNLP 2019

Summarizing multiple metrics in one score
remains an open problem.

Richard Yuanzhe Pang & Kevin Gimpel

Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer.

In Proceedings of 3rd NGT Workshop (@EMNLP) 2019

Remi Mir, Bjarke Felbo, Nick Obradovich & Iyad Rahwan

Evaluating style transfer for text.

In Proceedings of NAACL 2019

Models exhibit trafeoffs between evaluation dimensions

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov & Alexey Tikhonov

Style transfer and paraphrase: Looking for a sensible semantic similarity metric.

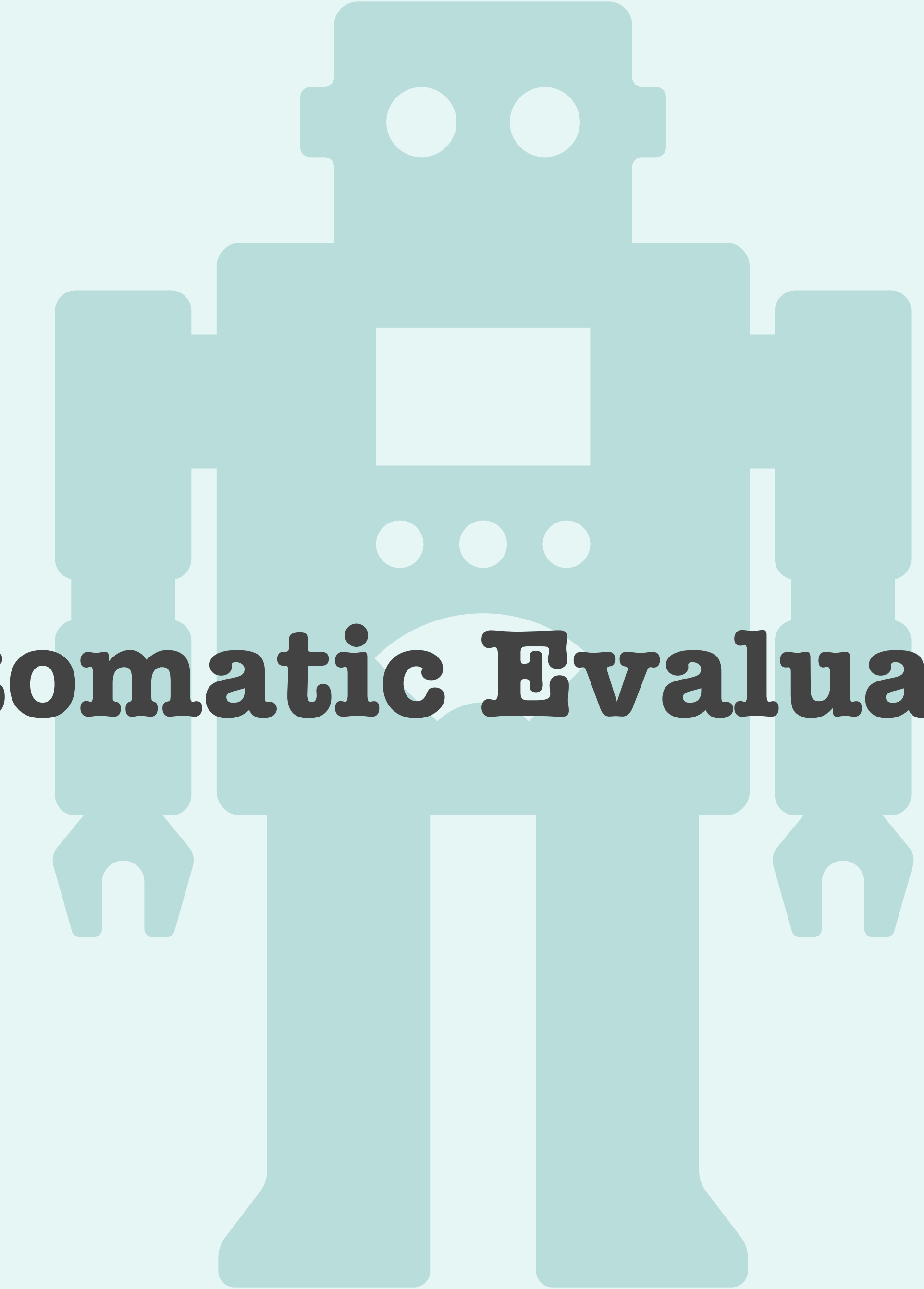
In Proceedings of AAAI 2021

Meta-evaluation on semantic similarity shows none of the
widely used metrics is close enough to human ratings

What are the best EVALUATION practices for Style Transfer?

What are the best EVALUATION practices for Style Transfer?

Human Evaluation



Automatic Evaluation

What are the best EVALUATION practices for Style Transfer?

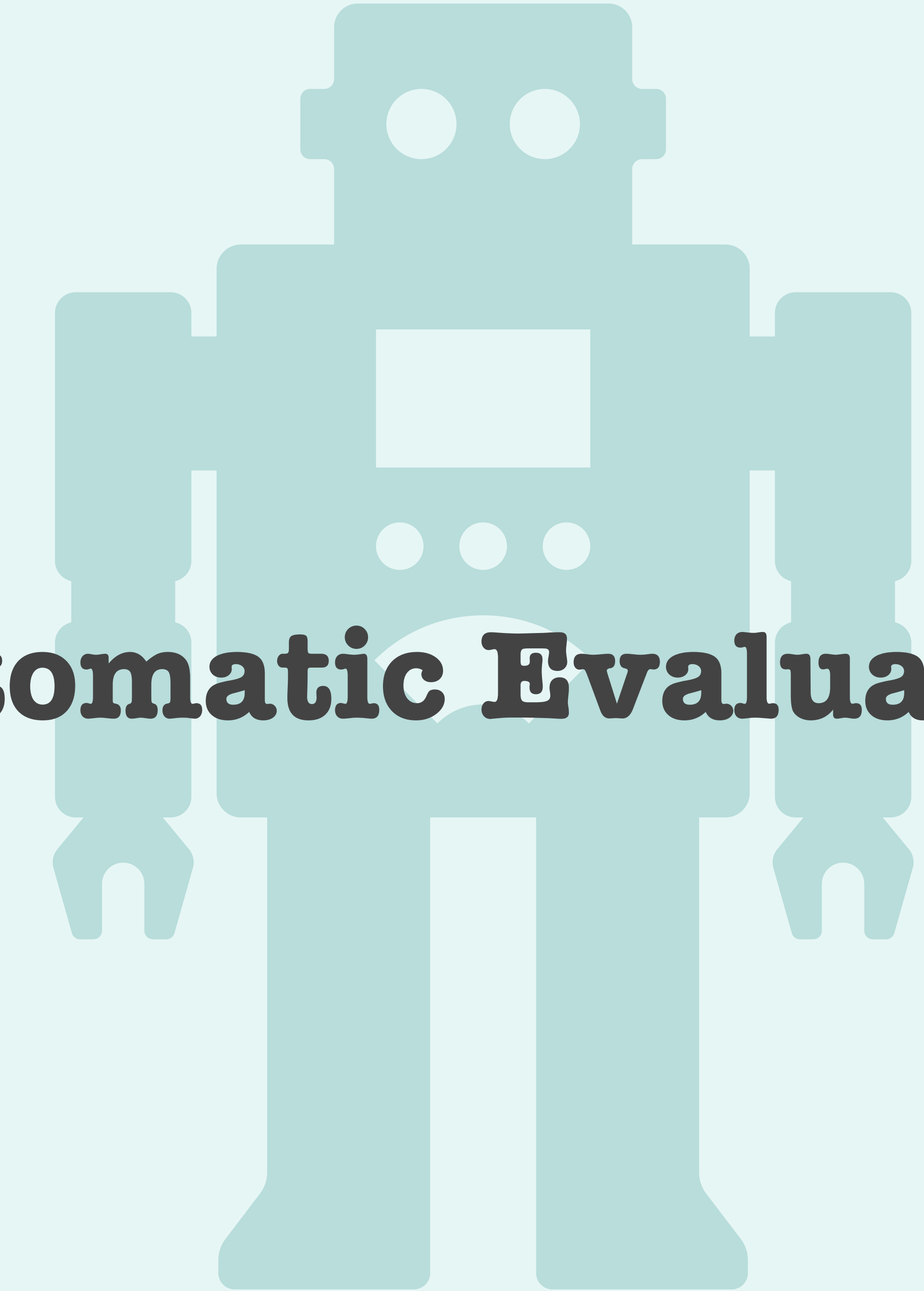
Human Evaluation

Automatic Evaluation

Outline:

Review current practices | Identify **limitations** | Provide **recommendations**

What are the best EVALUATION practices for Style Transfer?

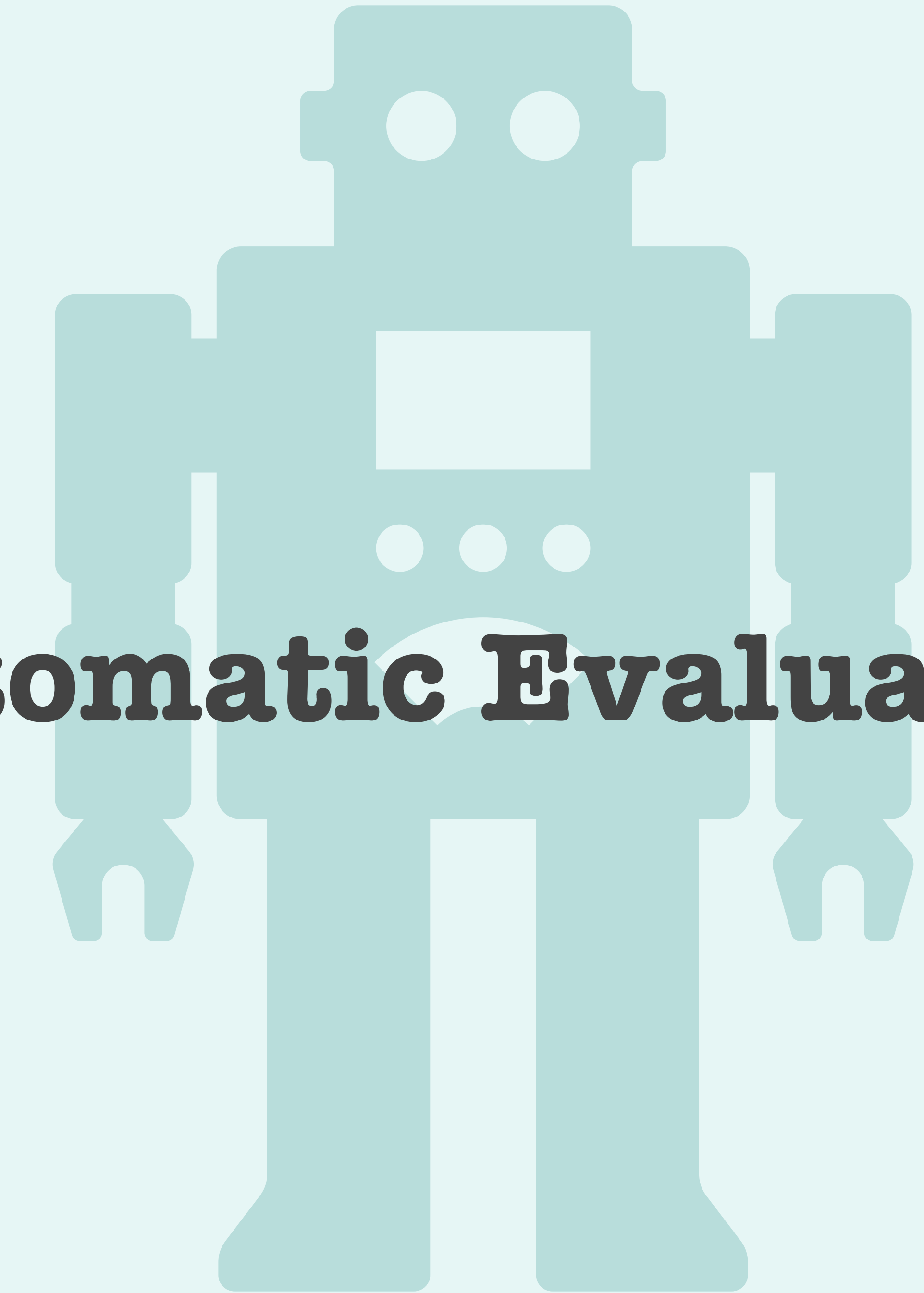


Outline:

Review current practices	Identify **limitations**	Provide **recommendations**

What are the best EVALUATION practices for Style Transfer?

Human Evaluation



Automatic Evaluation

Outline:

Review current practices	Identify **limitations**	Provide **recommendations**

What are the best EVALUATION practices for Style Transfer?

Human Evaluation

Automatic Evaluation

Outline:

Review current practices	Identify **limitations**	Provide **recommendations**

What are the best **EVALUATION** practices for Style Transfer?

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault & Marine Carpuat. 2021
A Review of **Human** Evaluation for Style Transfer.
In Proceedings of the First Workshop on Generation Evaluation and Metrics (GEM) at ACL.

A Structured Review of human EVALUATION for Style Transfer

A Structured Review of human EVALUATION for Style Transfer

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechotomova & Rada Mihalcea. 2021
Deep Learning for Text Style Transfer: A Survey.

fuzhenxin / Style-Transfer-in-Text Public

- 97 style transfer papers
- 86 from NLP & AI top-tier venues
- 11 pre-prints

as of March 2021

A Structured Review of human EVALUATION for Style Transfer

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechotomova & Rada Mihalcea. 2021
Deep Learning for Text Style Transfer: A Survey.

fuzhenxin / Style-Transfer-in-Text Public

- 97 style transfer papers
- 86 from NLP & AI top-tier venues
- 11 pre-prints

as of March 2021

GLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator's details
- Annotator's compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations' availability

A Structured Review of human EVALUATION for Style Transfer

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechotomova & Rada Mihalcea. 2021
Deep Learning for Text Style Transfer: A Survey.

fuzhenxin / Style-Transfer-in-Text Public

- 97 style transfer papers
- 86 from NLP & AI top-tier venues
- 11 pre-prints

as of March 2021

GLOBAL CRITERIA

- Task(s)
- Presence of human annotation
- Annotator's details
- Annotator's compensation
- Quality control
- Agreement statistics
- Evaluated systems
- Size of evaluated instance set
- Size of annotations per instance
- Sampling method
- Annotations' availability

DIMENSION-SPECIFIC CRITERIA

Fluency
Style
Meaning

- Presence of human evaluation
- Quality criterion name
- Form of response elicitation
- Details on collected responses
- Size of rating instrument

Howcroft et al. 2020
Twenty Years of Confusion in Human Evaluation:
NLG Needs Evaluation Sheets and Standardized Definitions
In Proceedings of the 13th International Conference
on Natural Language Generation.

How often do we rely on human EVALUATION?

Yes

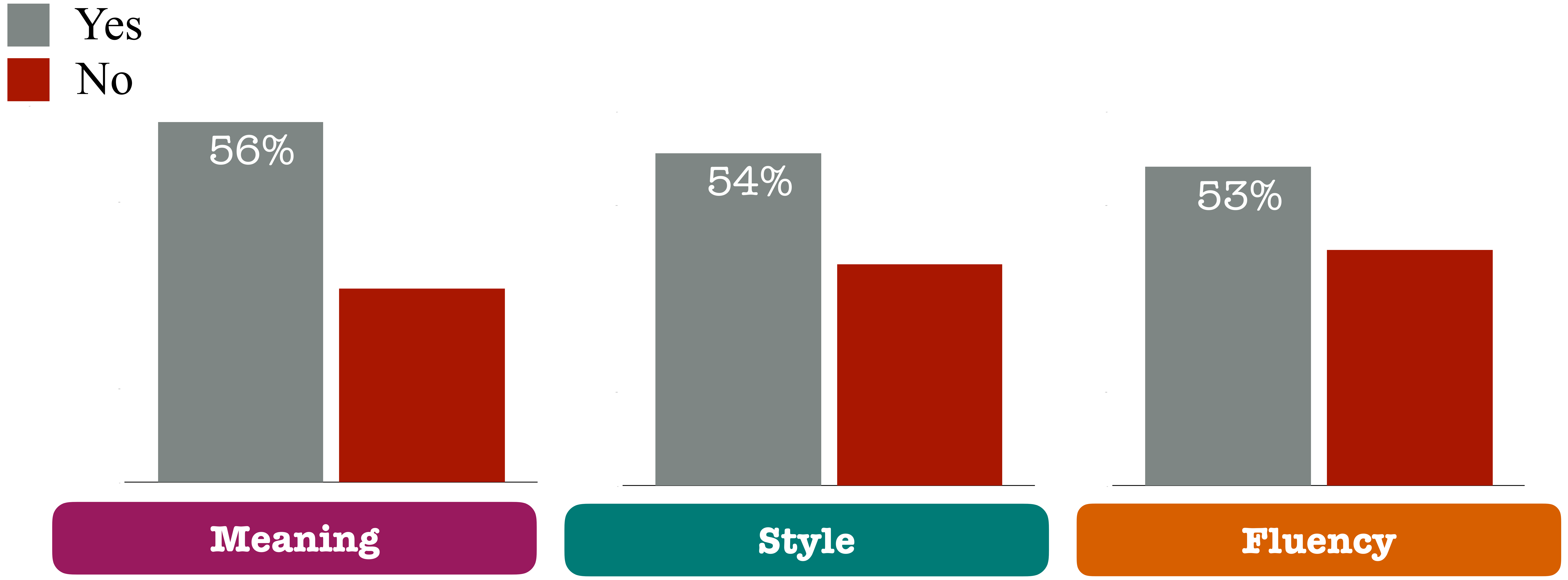
No

Meaning

Style

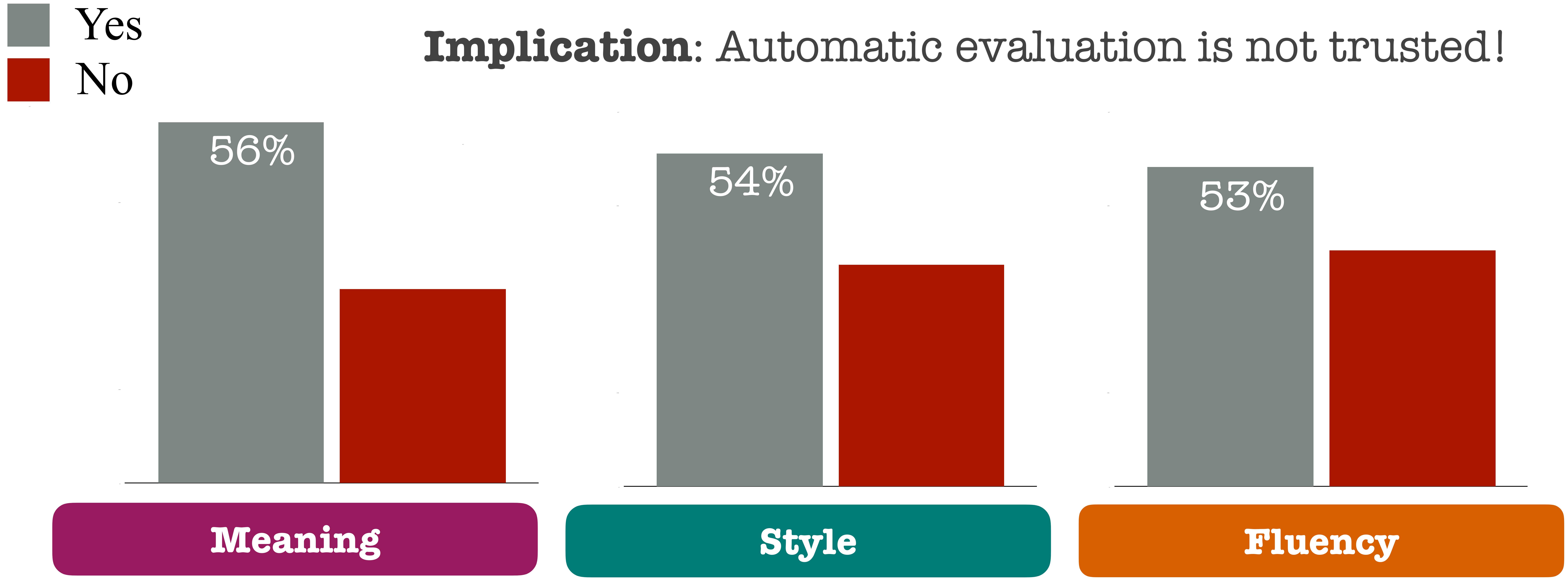
Fluency

> 50% of papers resort to human EVALUATION



> 50% of papers resort to human EVALUATION

Implication: Automatic evaluation is not trusted!



What are the limitations of current human EVALUATION practices for Style Transfer?

- **Underspecification**
  human annotation design attributes are missing
- **Availability**
  do not release the human ratings
- **Reliability**
  do not give details that can help assess their quality
- **Lack of standardization**
  inconsistent annotation protocols across papers

What are the limitations of current human EVALUATION practices for Style Transfer?

Underspecification
human annotation design attributes are missing

Availability
do not release the human ratings

Reliability
do not give details that can help assess their quality

Lack of standardization
inconsistent annotation protocols across papers

What are the limitations of current human EVALUATION practices for Style Transfer?

Underspecification
human annotation design attributes are missing

Availability
do not release the human ratings

Reliability
do not give details that can help assess their quality

Lack of standardization
inconsistent annotation protocols across papers

What are the limitations of current human EVALUATION practices for Style Transfer?

Underspecification
human annotation design attributes are missing

Availability
do not release the human ratings

Reliability
do not give details that can help assess their quality

Lack of standardization
inconsistent annotation protocols across papers

What are the limitations of current human EVALUATION practices for Style Transfer?

Underspecification
human annotation design attributes are missing

Availability
do not release the human ratings

Reliability
do not give details that can help assess their quality

Lack of standardization
inconsistent annotation protocols across papers

What are the best EVALUATION practices for Style Transfer?

Human Evaluation

- Describe evaluation protocols
- Release annotations
- Standardize evaluation protocols

What are the best EVALUATION practices for Style Transfer?

Eleftheria Briakou, Sweta Agrawal,
Joel Tetreault & Marine Carpuat. 2021
Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer.
In Proceedings of the 2021 Conference on Empirical Methods in Natural Language (EMNLP) Processing.

For formality transfer: Data is more standardized, but **EVALUATION** is not

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	✗

What are the limitations of current automatic EVALUATION practices for Style Transfer?

Style
- Linear regressor
- CCN classifier
- CCN classifier
- CCN classifier
- LSTM classifier
- CCN classifier
- LSTM classifier
- RoBerta classifier
- CCN classifier
- GRU classifier
- BERT classifier
- FASTTEXT classifer
- CNN classifier
- CNN classifier
- CNN classifier
- GRU classifier
- RoBerta classifier
- CNN classifier
- BERT regressor

Meaning
- CCN classifier
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- Embedding Similarity
- Embedding Similarity
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- reference-BLEU
- reference-BLEU
- self-BLEU

Fluency
- Linear regressor
- RNN-LM perplexity
- RoBerta-LM perplexity
- Linear regressor
- KenLM perplexity
- GPT-LM perplexity
- LSTM-LM perplexity
- GPT-LM perplexity
- GPT-LM perplexity
- KenLM perplexity

What are the limitations of current automatic EVALUATION practices for Style Transfer?

Lack of standardized metrics

Style
- Linear regressor
- CCN classifier
- CCN classifier
- CCN classifier
- LSTM classifier
- CCN classifier
- LSTM classifier
- RoBerta classifier
- CCN classifier
- GRU classifier
- BERT classifier
- FASTTEXT classifier
- CNN classifier
- CNN classifier
- CNN classifier
- GRU classifier
- RoBerta classifier
- CNN classifier
- BERT regressor

Meaning
- CCN classifier
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- Embedding Similarity
- Embedding Similarity
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- reference-BLEU
- reference-BLEU
- self-BLEU

Fluency
- Linear regressor
- RNN-LM perplexity
- RoBerta-LM perplexity
- Linear regressor
- KenLM perplexity
- GPT-LM perplexity
- LSTM-LM perplexity
- GPT-LM perplexity
- GPT-LM perplexity
- KenLM perplexity

What are the limitations of current automatic EVALUATION practices for Style Transfer?

- Lack of standardized metrics
- Complemented by human evaluation

Style 11/19

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

Meaning 11/17

CCN classifier
reference-BLEU
reference-BLEU
reference-BLEU
BERT classifier
Embedding Similarity
Embedding Similarity
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
BERT classifier
reference-BLEU
reference-BLEU
self-BLEU

Fluency 6/10

Linear regressor
RNN-LM perplexity
RoBerta-LM perplexity
Linear regressor
KenLM perplexity
GPT-LM perplexity
LSTM-LM perplexity
GPT-LM perplexity
GPT-LM perplexity
KenLM perplexity

What are the limitations of current automatic EVALUATION practices for Style Transfer?

- Lack of standardized metrics
- Complemented by human evaluation
- Lack of agreement with human judgments

Style

- ~~Linear regressor~~
- CCN classifier
- ~~CCN classifier~~
- ~~CCN classifier~~
- LSTM classifier
- CCN classifier
- LSTM classifier
- ~~RoBerta classifier~~
- ~~CCN classifier~~
- ~~GRU classifier~~
- BERT classifier
- ~~FASTTEXT classifier~~
- CNN classifier
- CNN classifier
- CNN classifier
- GRU classifier
- RoBerta classifier
- CNN classifier
- ~~BERT regressor~~

Meaning

- CCN classifier
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- Embedding Similarity
- Embedding Similarity
- ~~reference-BLEU~~
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- reference-BLEU
- BERT classifier
- reference-BLEU
- reference-BLEU
- self-BLEU

Fluency

- ~~Linear regressor~~
- RNN-LM perplexity
- RoBerta-LM perplexity
- ~~Linear regressor~~
- ~~KenLM perplexity~~
- GPT-LM perplexity
- LSTM-LM perplexity
- GPT-LM perplexity
- GPT-LM perplexity
- ~~KenLM perplexity~~

What are the limitations of current automatic EVALUATION practices for Style Transfer?

- Lack of standardized metrics
- Complemented by human evaluation
- Lack of agreement with human judgments
- Lack of portability to multiple languages

Style

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

Meaning

CCN classifier EN
reference-BLEU
reference-BLEU
reference-BLEU
BERT classifier EN
Embedding Similarity
Embedding Similarity
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
reference-BLEU
BERT classifier
reference-BLEU
reference-BLEU
self-BLEU

Fluency

Linear regressor EN
RNN-LM perplexity
RoBerta-LM perplexity
Linear regressor EN
KenLM perplexity
GPT-LM perplexity
LSTM-LM perplexity
GPT-LM perplexity
GPT-LM perplexity
KenLM perplexity

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Human Ratings

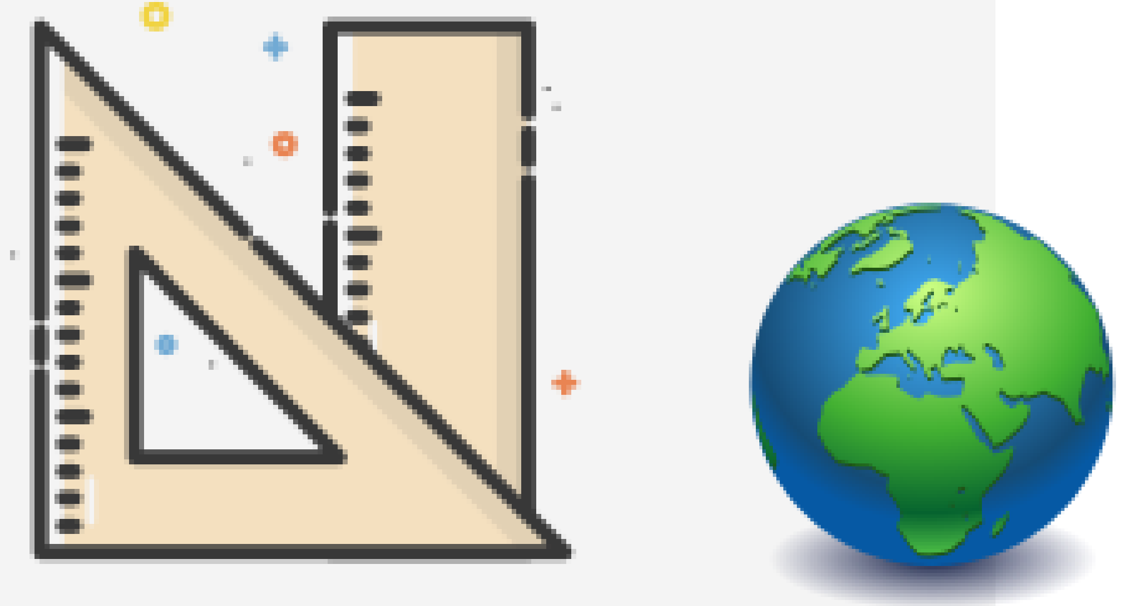
System Outputs

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

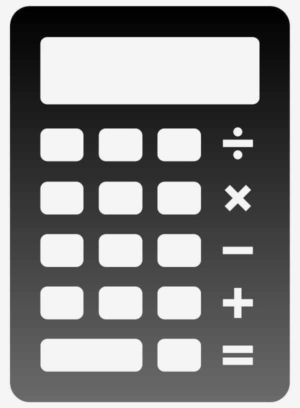
System Outputs

Human Ratings

Automatic Ratings



Correlation Analysis



Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Fluency

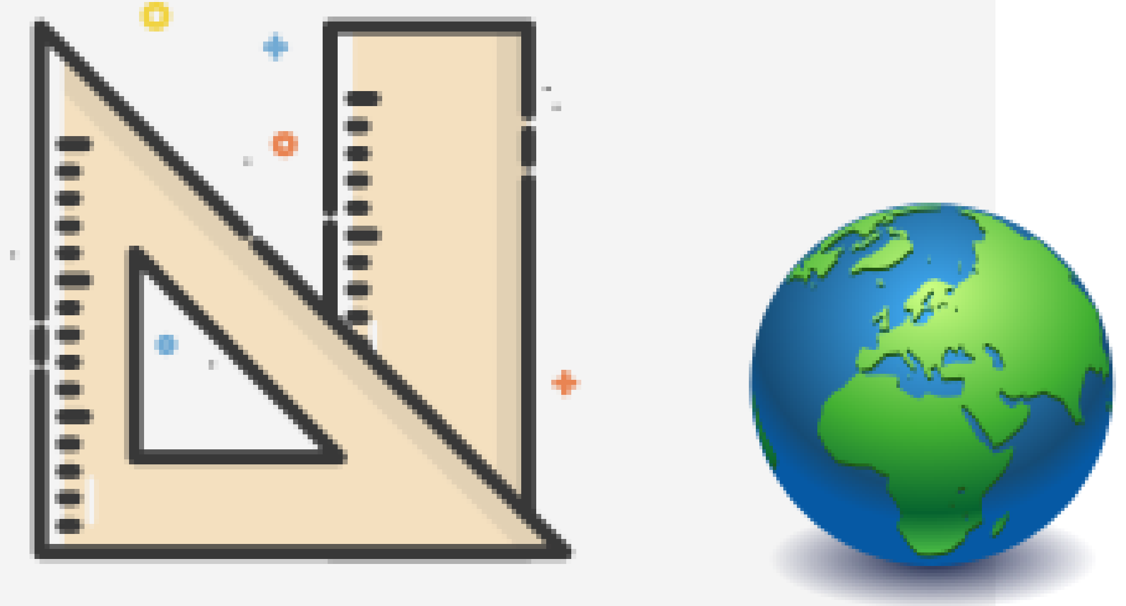
Style

Meaning

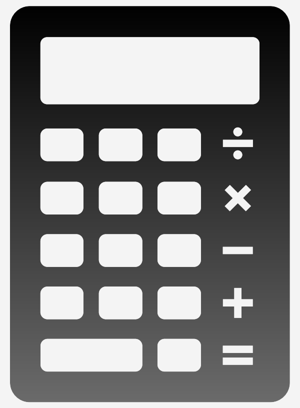
System Outputs

Human Ratings

Automatic Ratings



Correlation Analysis



Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Human Ratings collected **consistently** across dimensions & languages

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Human Ratings collected **consistently** across dimensions & languages

Sudha Rao & Joel Tetreault. 2018
Dear sir or madam, may I introduce the GYAFC corpus.
In Proceedings of NAACL-HLT.

English (EN)

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Human Ratings collected **consistently** across dimensions & languages

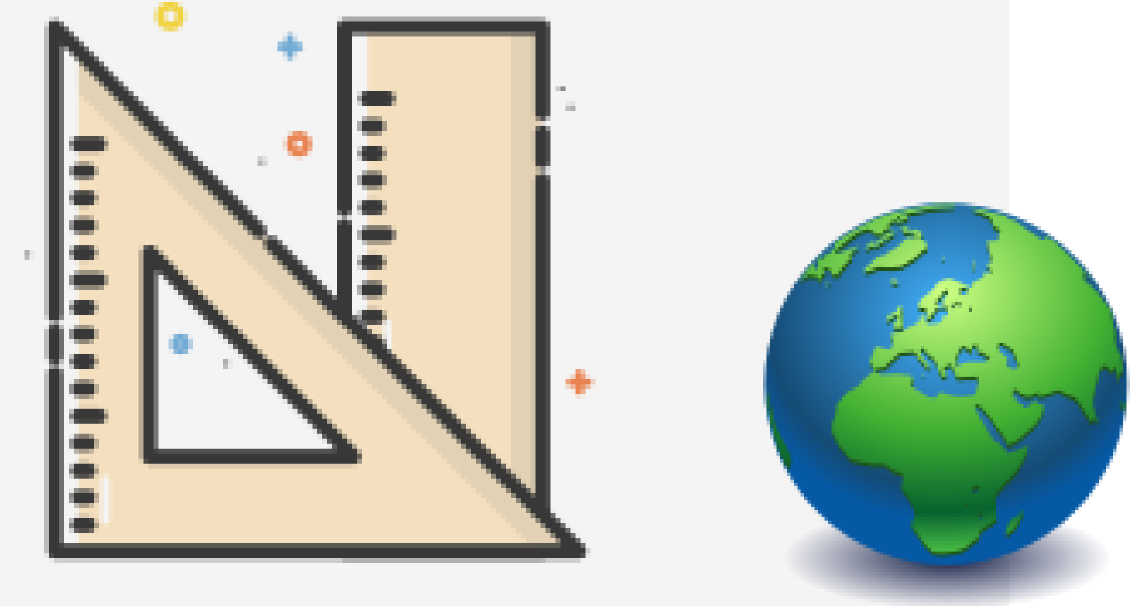
Sudha Rao & Joel Tetreault. 2018
Dear sir or madam, may I introduce the GYAFC corpus.
In Proceedings of NAACL-HLT.

Eleftheria Briakou, Di Lu, Ke Zhang, Joel Tetreault . 2021
Olá, Bonjour, Salve! XFORMAL: A Benchmark for Multilingual Formality Style Transfer.
In Proceedings of NAACL-HLT.

- English (EN)
- Brazilian-Portuguese (BR-PT)
- Italian (IT)
- French (FR)

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Automatic Ratings



Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Style

Approach: Supervised

Models: multilingual pre-trained LMs

Cross-lingual Transfer:

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

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Style

Approach: Supervised

Models: multilingual pre-trained LMs

Cross-lingual Transfer:

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

Meaning

Approach: Supervised; Unsupervised; String-based

Models: multilingual pre-trained LMs; embedding based

following meta-evaluation of:

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, Alexey Tikhonov. 2021 Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric. In Proceedings of AAAI.

Empirical EVALUATION of Automatic Metrics For Formality Style Transfer Evaluation

Style

Approach: Supervised

Models: multilingual pre-trained LMs

Cross-lingual Transfer:

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

Meaning

Approach: Supervised; Unsupervised; String-based

Models: multilingual pre-trained LMs; embedding based

following meta-evaluation of:

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, Alexey Tikhonov. 2021 Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric In Proceedings of AAAI.

Fluency

Approach: Unsupervised

Models: language models

- kenLM
- mBERT
- XLM-R

Metrics: perplexity

Best Practice for Formality Evaluation ?

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

Best Practice for Formality Evaluation ?

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

Most common practice:

- Classification: Formal vs. Informal
- Evaluated on human written testbed

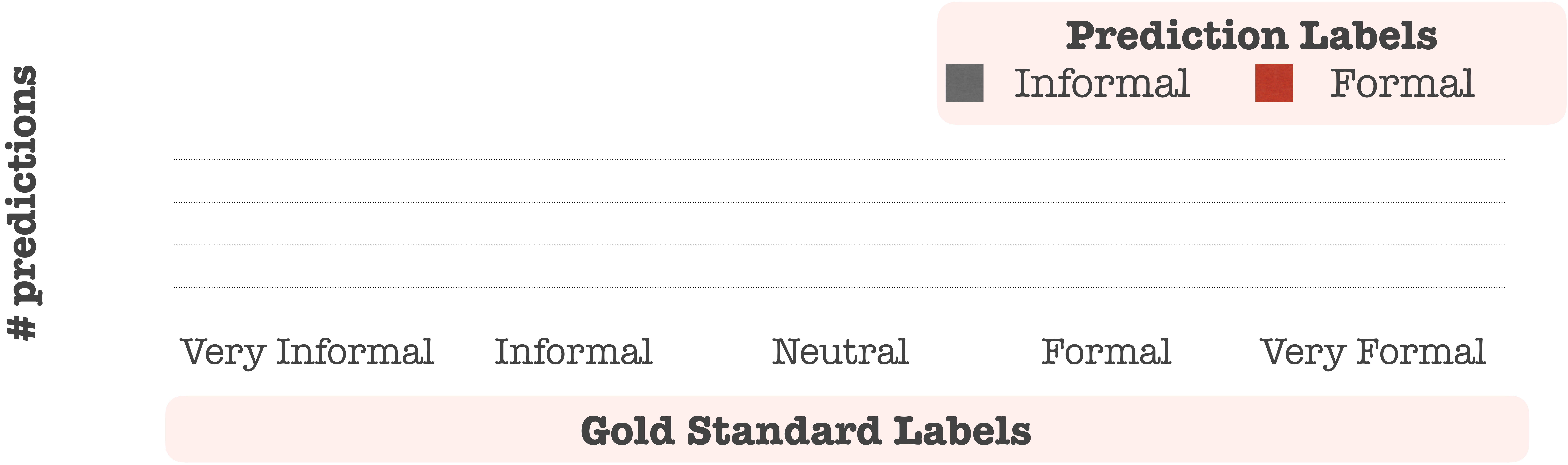
Best Practice for Formality Evaluation ?

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

Most common practice:

- Classification: Formal vs. Informal
- Evaluated on human written testbed
- How does it perform on system outputs?

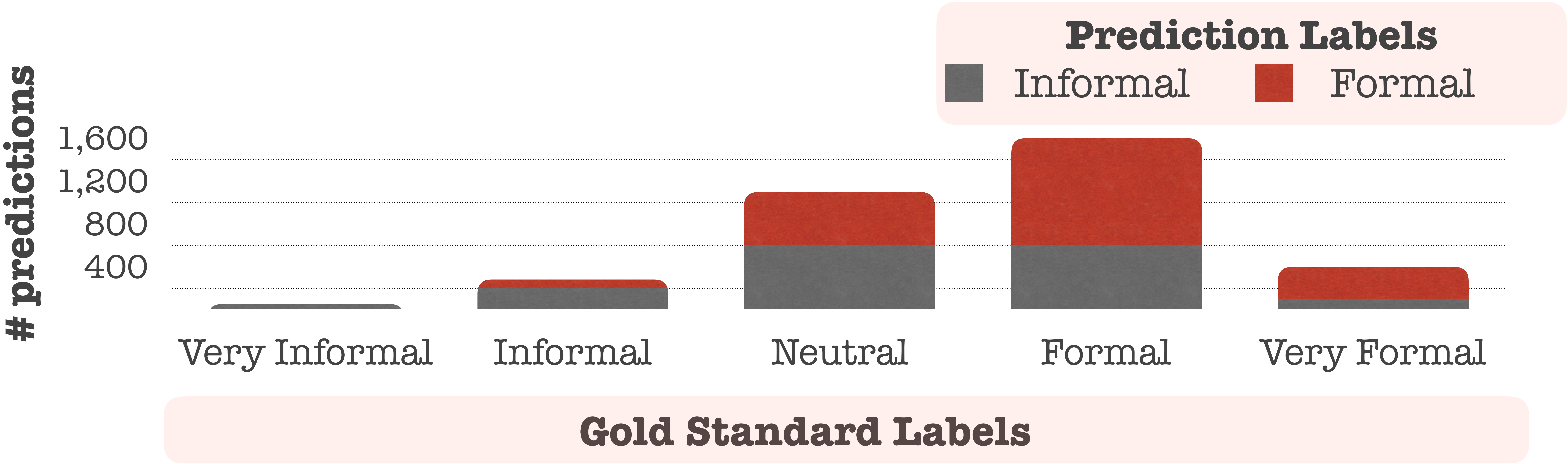
Binary Formality Classifiers



Per bin analysis: Human ratings are given at a fine-grained level!

Binary Formality Classifiers

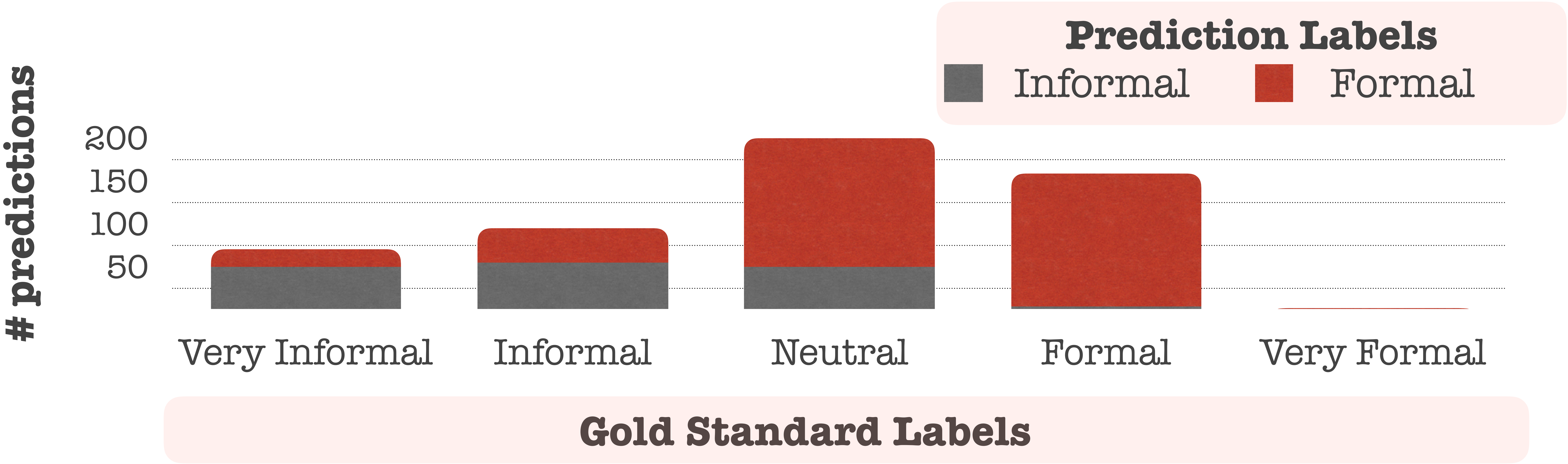
Lack sensitivity to different formality levels



English

Binary Formality Classifiers

Are biased towards the formal class



French

Best Practice for Formality Evaluation ?

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

Most common practice:

- Classification: Formal vs. Informal
- Evaluated on human written testbed
- How does it perform on system outputs?

Proposed practice:

- Regression: Fine-grained formality levels
- Evaluated on system outputs written testbed

Best Practice for Formality Evaluation ?

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

Most common practice:

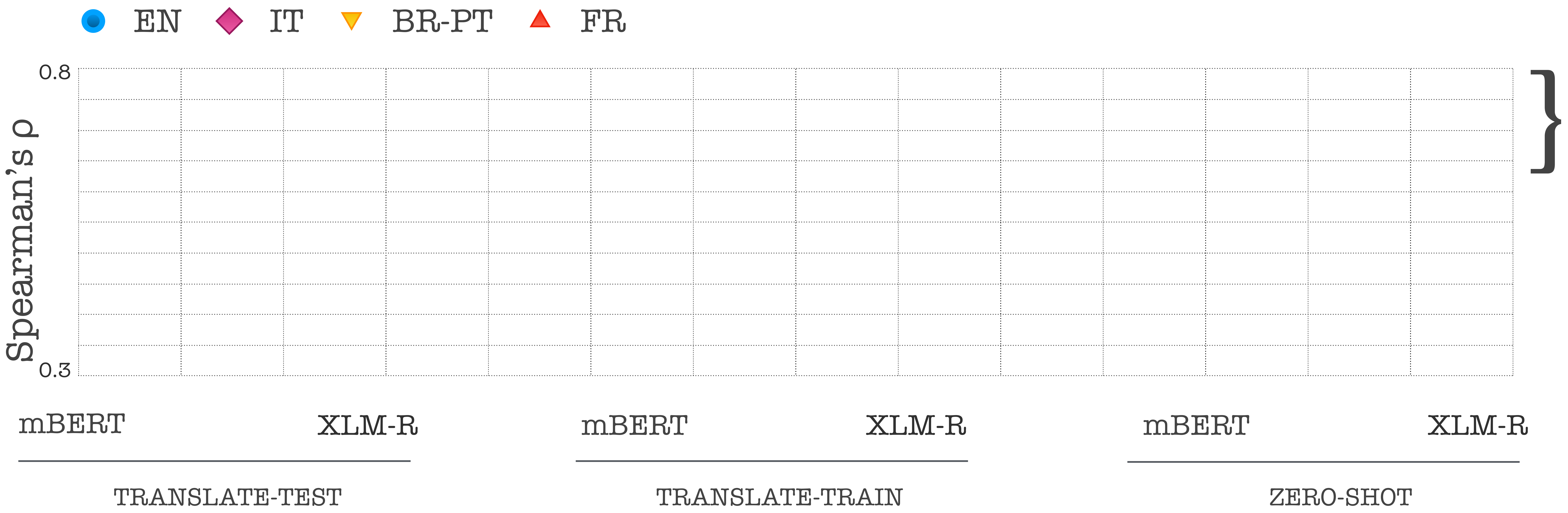
- Classification: Formal vs. Informal
- Evaluated on human written testbed
- How does it perform on system outputs?

Proposed practice:

- Regression: Fine-grained formality levels
- Evaluated on system outputs written testbed
- How does it perform on system outputs?

Best Practice for Formality Evaluation

XLM-R yields best correlations across languages



Best Practice for Formality Evaluation

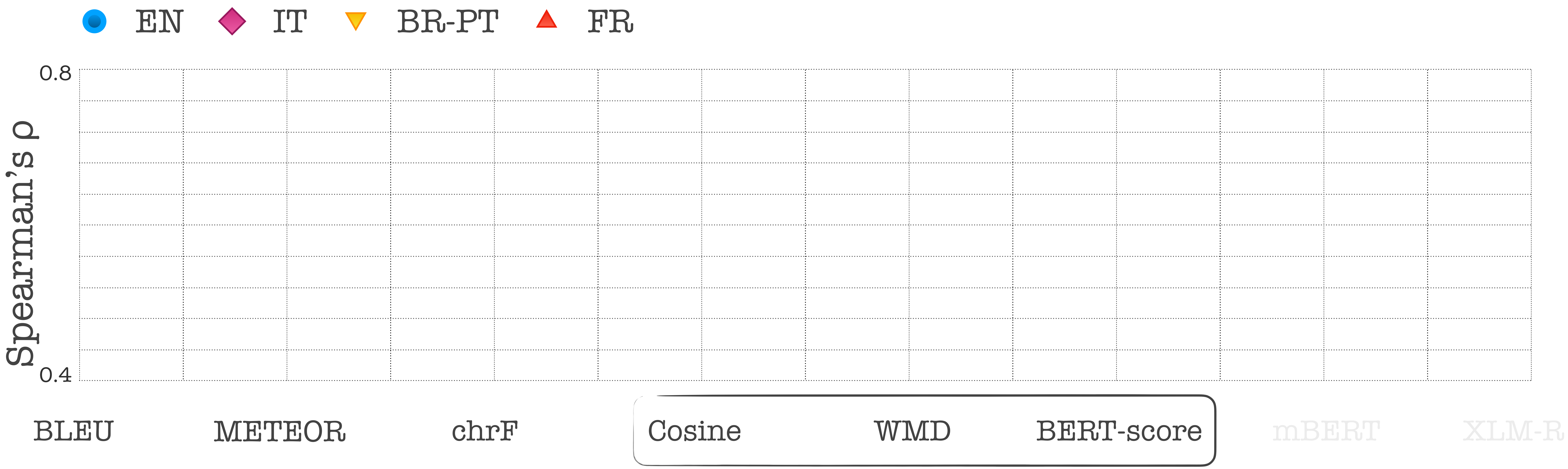
XLM-R yields best correlations across languages

Best Practice for Meaning Evaluation

Best Practice for Meaning Evaluation

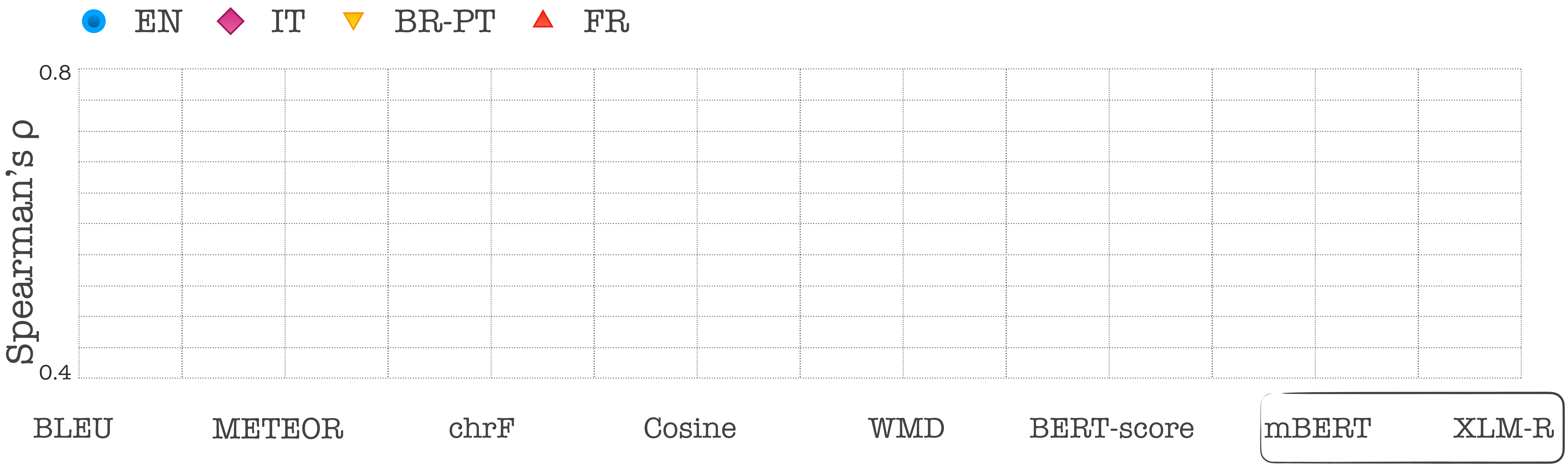
String-based: compare system output with system input

Best Practice for Meaning Evaluation



Unsupervised: based on pre-trained embeddings

Best Practice for Meaning Evaluation



Supervised: trained on Semantic Textual Similarity data

Best Practice for Meaning Evaluation

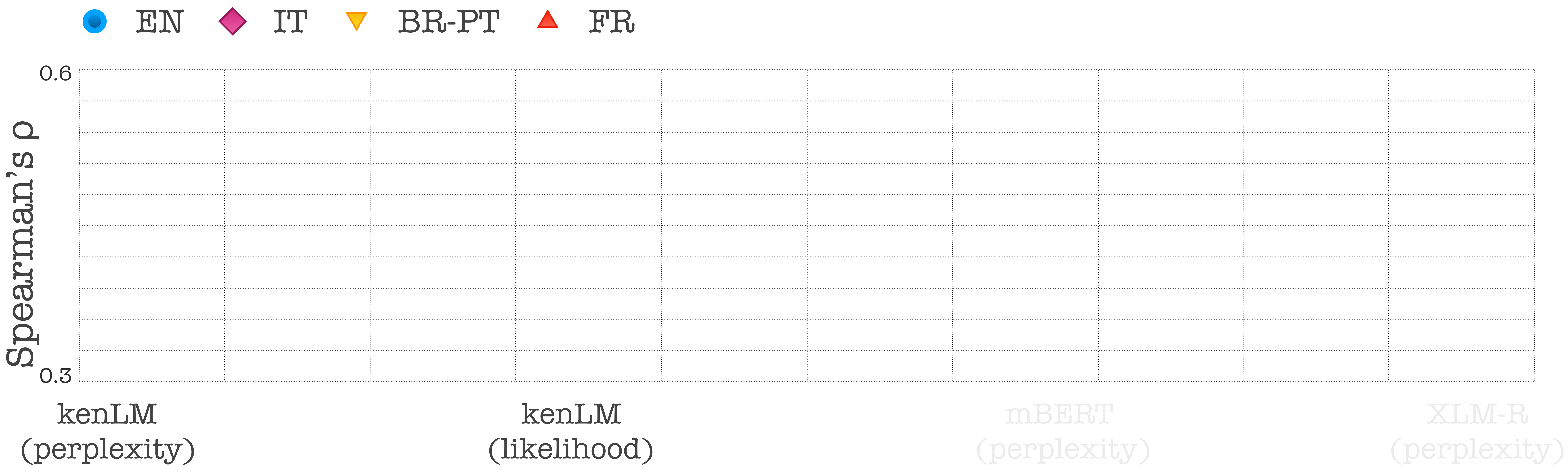
All metrics yield above moderate correlations

Best Practice for Meaning Evaluation

chrF yields best correlations across languages

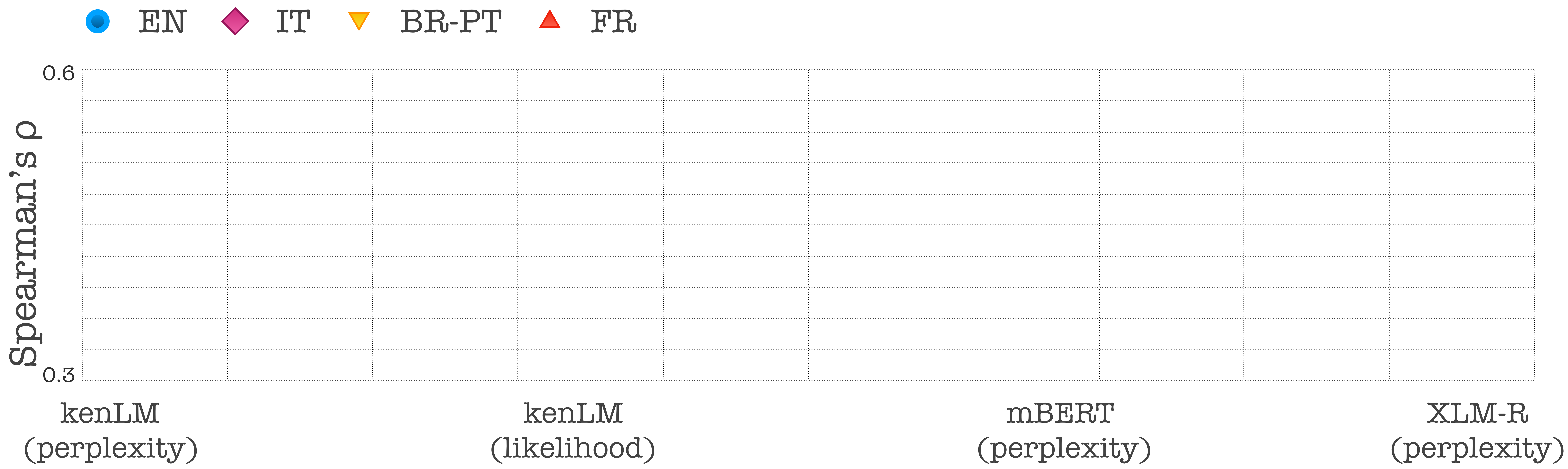
Best Practice for Fluency Evaluation

Best Practice for Fluency Evaluation



N-gram based LM: trained on monolingual texts

Best Practice for Fluency Evaluation



Pre-trained multilingual LMs: pseudo-perplexity

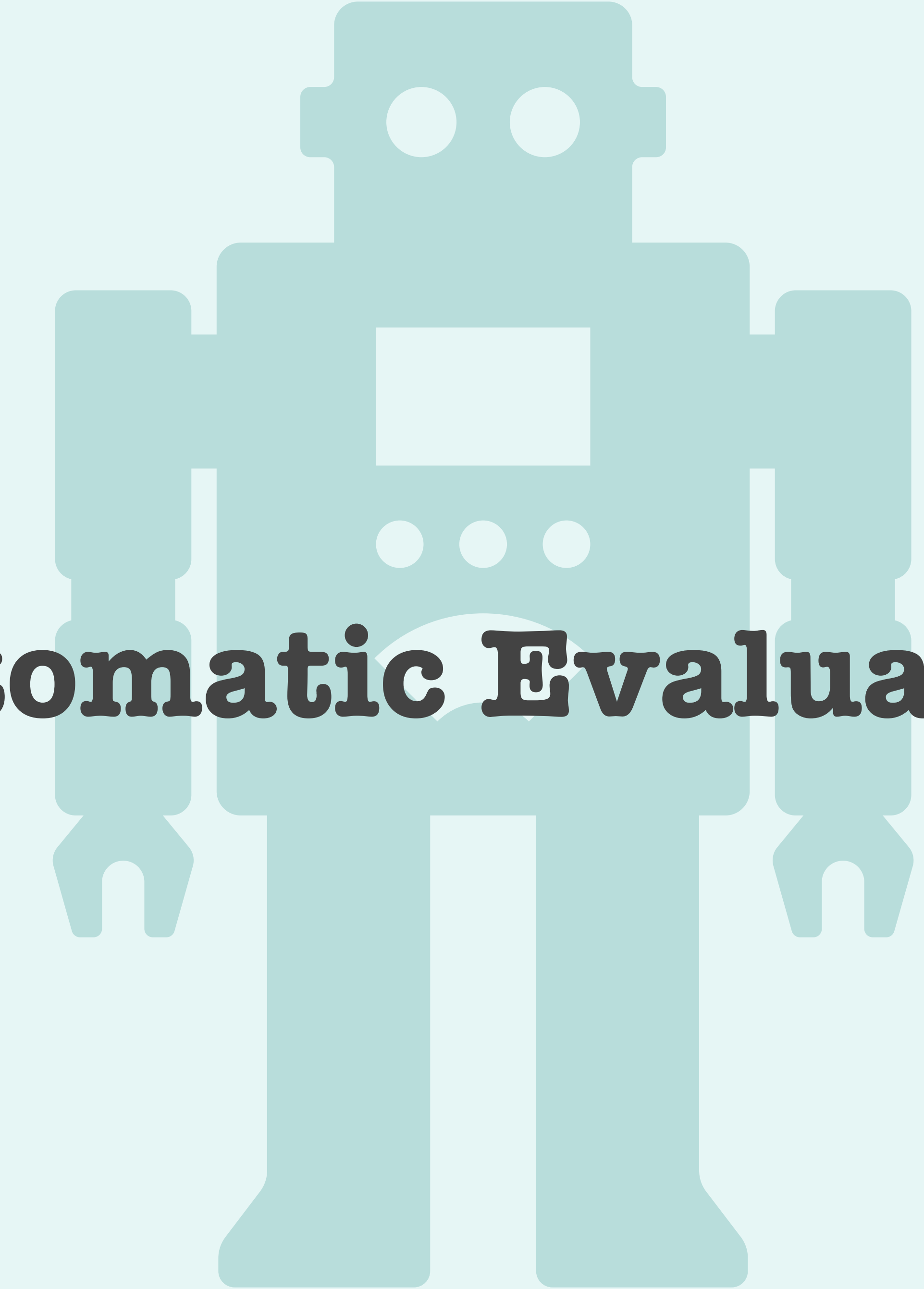
Best Practice for Fluency Evaluation

XLM-R yields best correlations across languages

What are the best EVALUATION practices for Style Transfer?

Human Evaluation

- ✓ Describe evaluation protocols
- ✓ Release annotations
- ✓ Standardize evaluation protocols



Automatic Evaluation

- ✓ **Style** : XLM-R regression (zero-shot)
- ✓ **Meaning** : chRF score with input references
- ✓ **Fluency** : XLM-R pseudo-perplexity

Open questions on Style Transfer EVALUATION

Open questions on Style Transfer EVALUATION

What can we learn from human disagreements about the nature of ST tasks?

How can we use annotation disagreements to model ST tasks?

How do different evaluation protocols bias the collected responses?

Open questions on Style Transfer **EVALUATION**

What can we learn from human disagreements about the nature of ST tasks?

How can we use annotation disagreements to model ST tasks?

How do different evaluation protocols bias the collected responses?

How do best practices generalize across more diverse languages?

How do best practices generalize across different definitions of style?

How should different metrics be aggregated in a single one?

QUESTIONS?

Eleftheria Briakou
email: ebriakou@umd.edu
twitter: @ebriakou

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault & Marine Carpuat. 2021
A Review of Human Evaluation for Style Transfer.
In Proceedings of the First Workshop on Generation Evaluation and Metrics (GEM) at ACL.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault & Marine Carpuat. 2021
Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer.
In Proceedings of the 2021 Conference on Empirical Methods in Natural Language (EMNLP) Processing.