Human Judgement as a Compass to Navigate Automatic Metrics for Formality Transfer

Huiyuan Lai Jiali Mao Antonio Toral Malvina Nissim
CLCG, University of Groningen / The Netherlands

HumEval2022

What is Text Style Transfer?

Generate a well-formed text in the target style while preserving the content/theme

What is Text Style Transfer?

Generate a well-formed text in the target style while preserving the content/theme

Source: *i like this screen, it's just the right size...*

What is Text Style Transfer?

Generate a well-formed text in the target style while preserving the content/theme

Sentiment Aspect

Source: *i like this screen, it's just the right size...*

Polarity Swap

Target: *i hate this screen, it is not the right size..*

What is Text Style Transfer?

Generate a well-formed text in the target style while preserving the content/theme

Formality Aspect

Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Style Transfer: Evaluation

Generate a well-formed text in the target style while preserving the content/theme

- ✓ In proper language, hence fluent and grammatical
- ✓ In the appropriate target style
- ✓ In a way such that the content/theme, is preserved

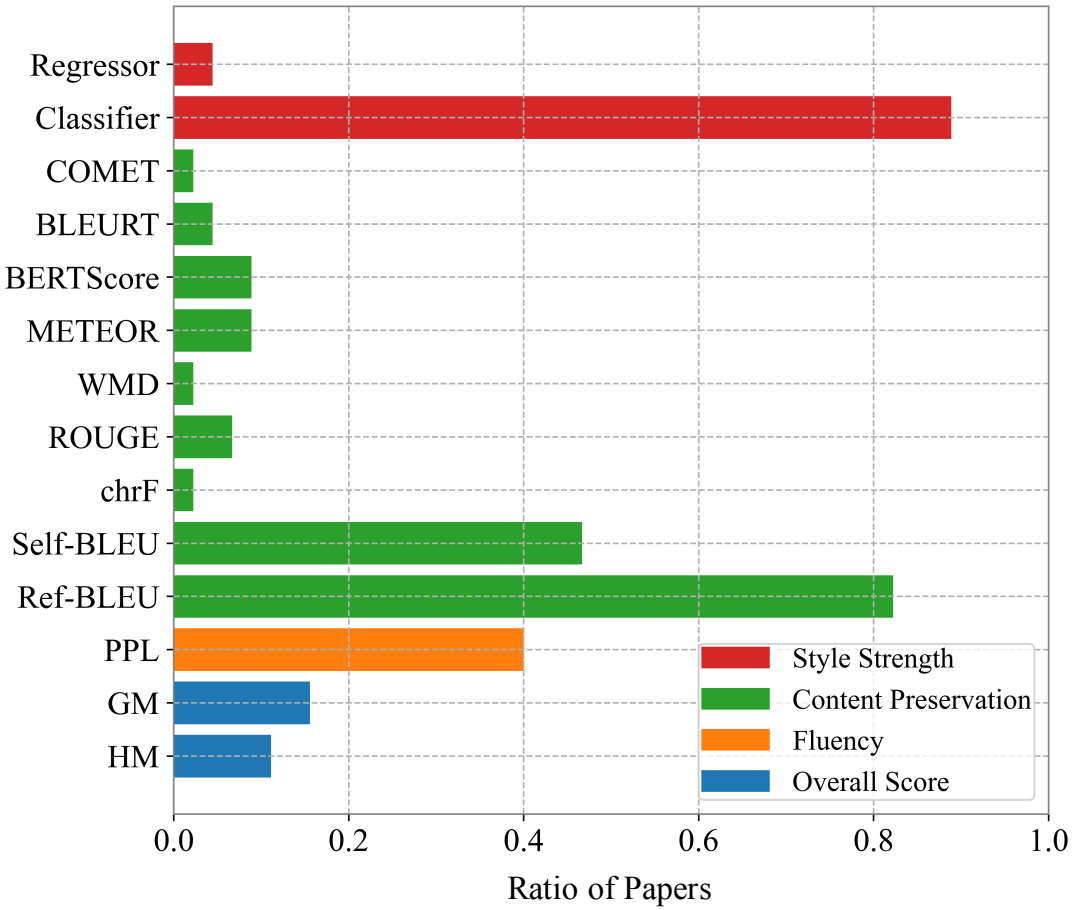
Style Transfer: Evaluation

Generate a well-formed text in the target style while preserving the content/theme

- ✓ In proper language, hence fluent and grammatical
- ✓ In the appropriate target style
- ✓ In a way such that the content/theme, is preserved

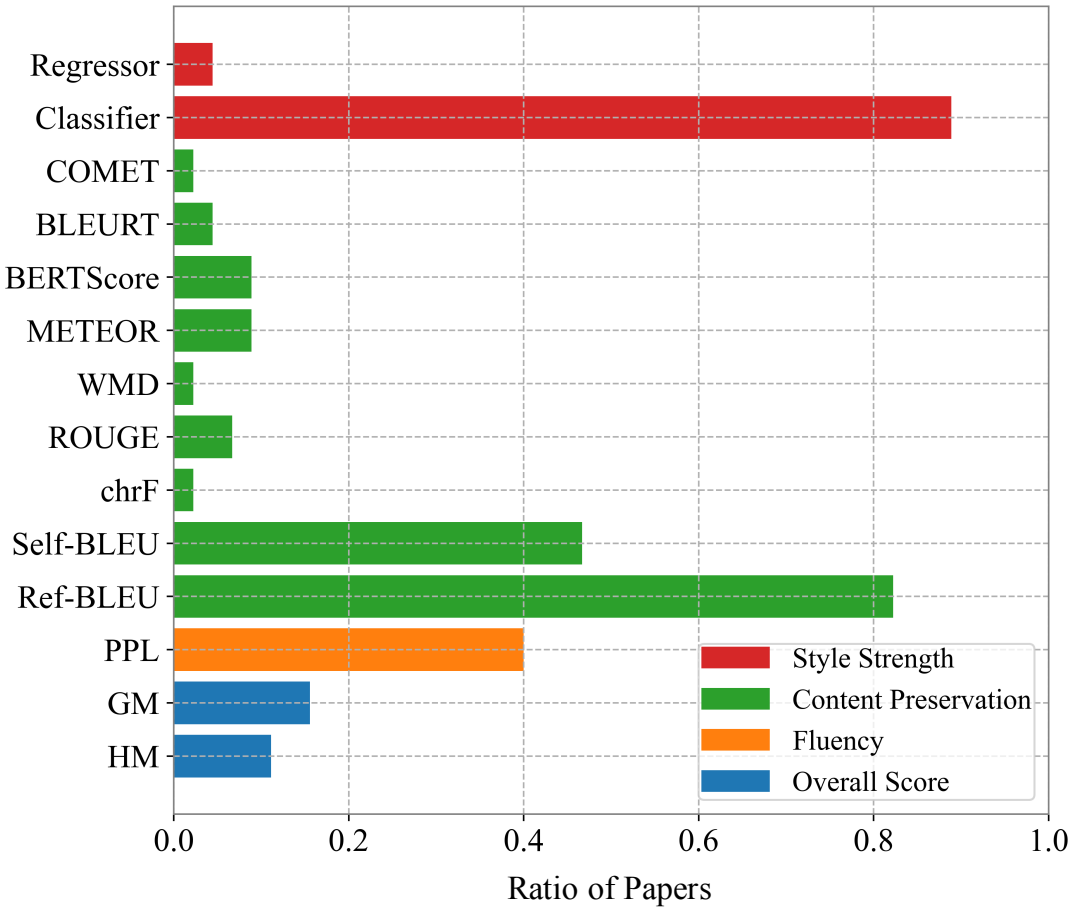
Fluency | **Style Strength** | **Content Preservation**

Style Transfer: Automatic Evaluation



Automatic evaluation metrics in 45 ACL Anthology papers focusing on style transfer.

Style Transfer: Evaluation

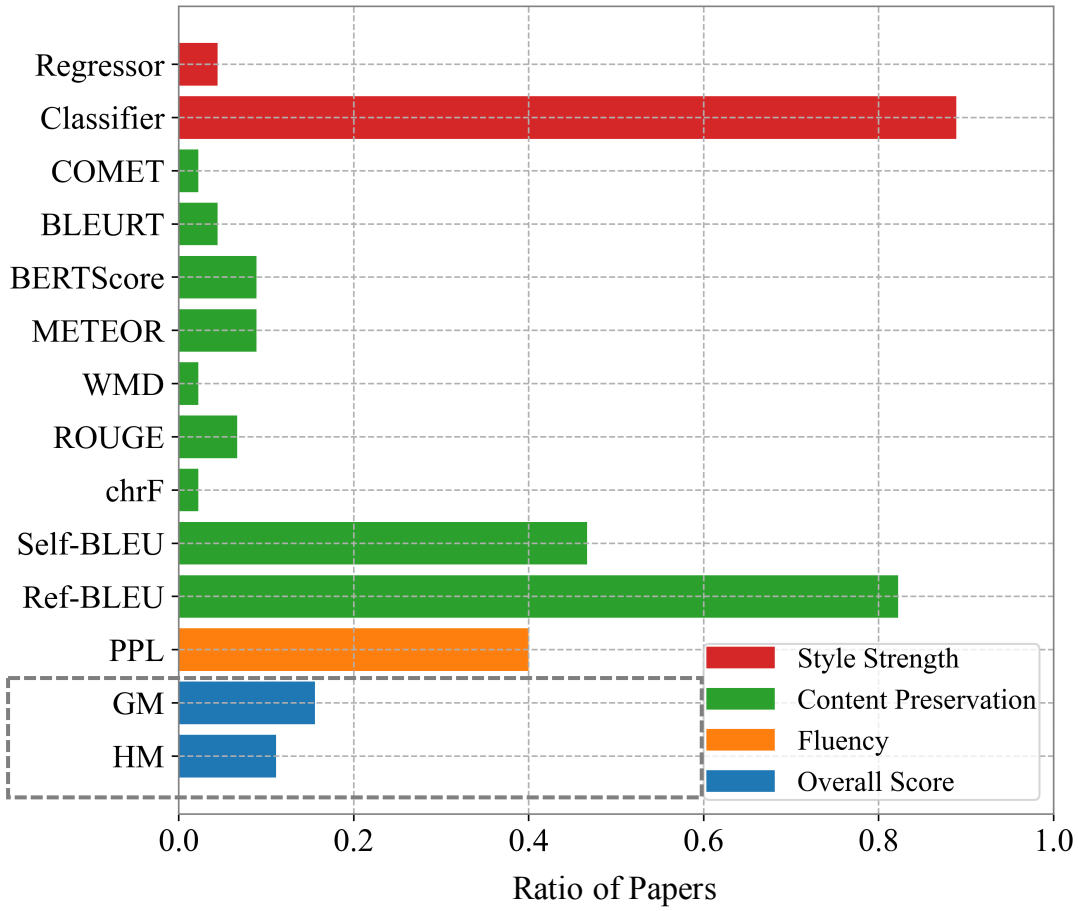


Style Strength

Content Preservation

Fluency

Style Transfer: Evaluation

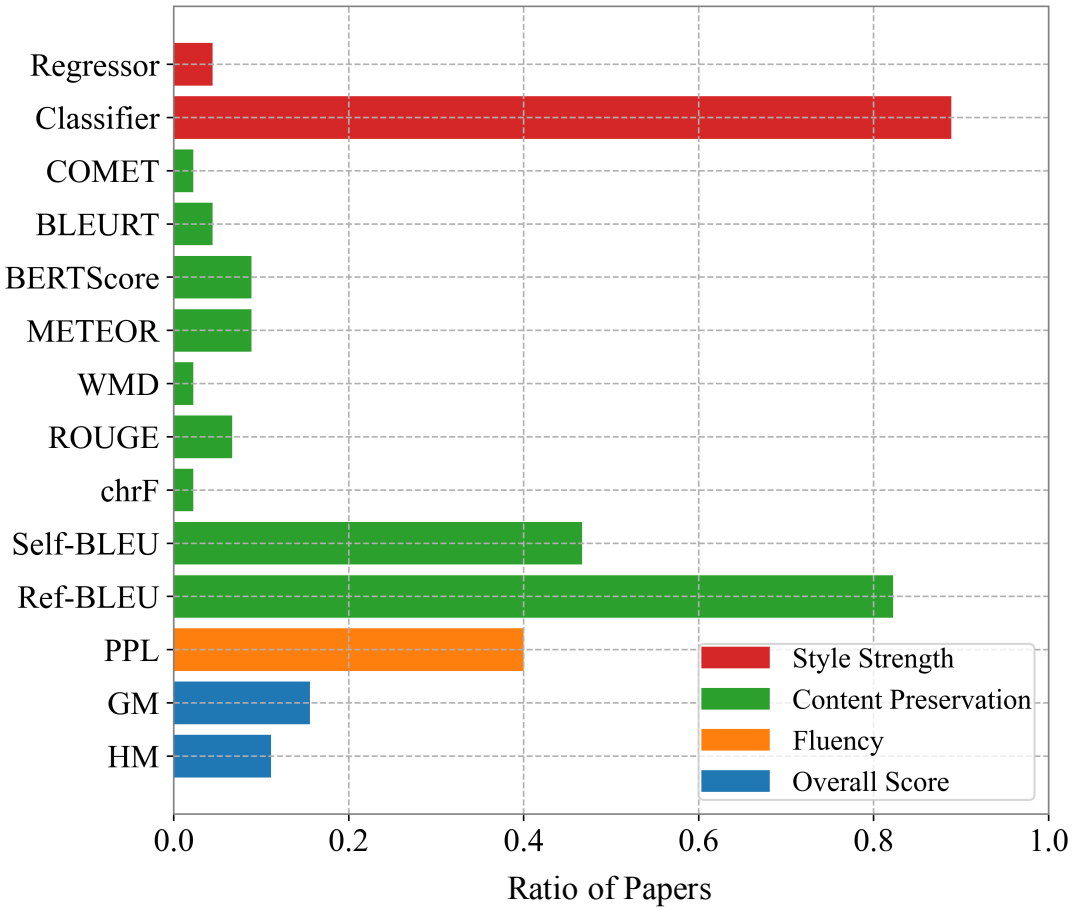


Style Strength

Content Preservation

Fluency

Style Transfer: Evaluation



- ✓ Correlations of human judgements and automatic metrics [Rao et al., 2018; Lai et al., 2021]
- ✓ Traditional metrics for polarity swap [Tikhonov et al., 2019; Mir et al., 2019]
- ✓ Content metrics in the context of formality transfer and paraphrasing [Yamshchikov et al., 2021]
- ✓ Automatic Metrics for multilingual formality transfer [Briakou et al., 2021]

Automatic Evaluation: Challenges

Automatic Evaluation: Challenges

TASKS

Tasks are conflated under the TST label while they are **not** exactly the same [Lai et al., 2021]

Polarity Swap

VS

Formality Transfer

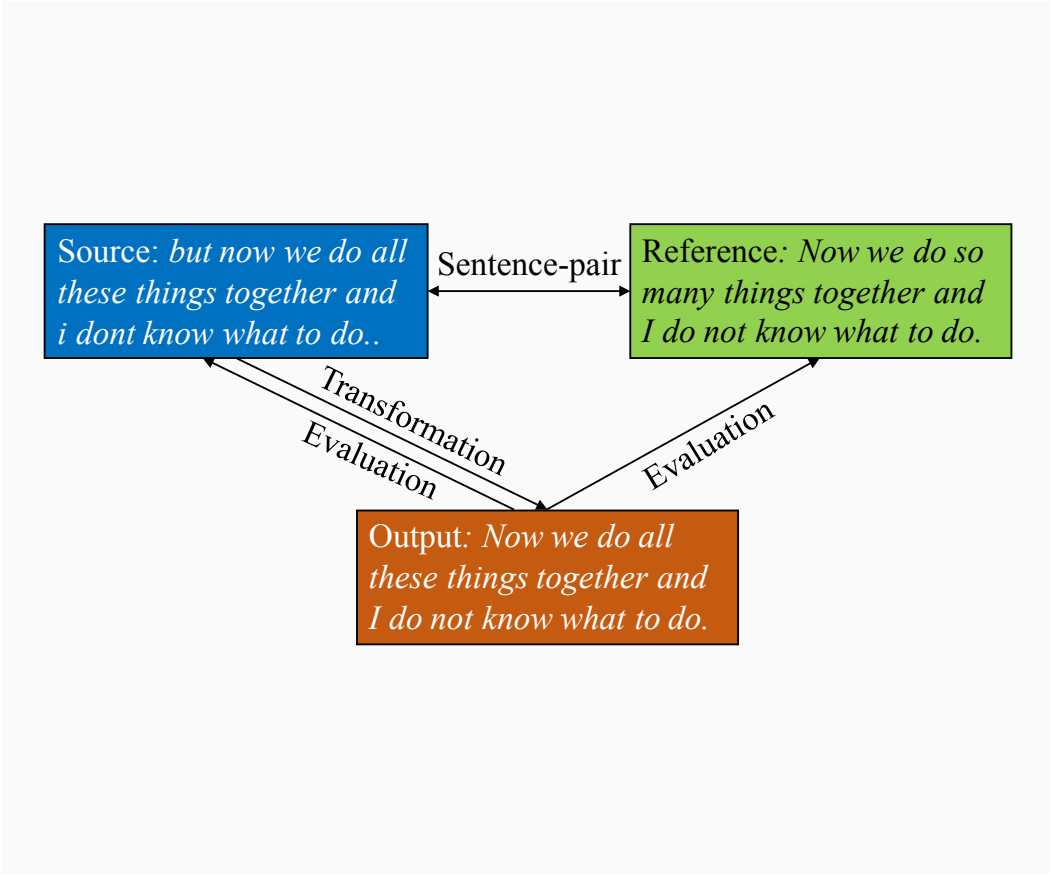
Automatic Evaluation: Challenges

TASKS

Tasks are conflated under the TST label while they are **not** exactly the same [Lai et al., 2021]

SETTINGS

The evaluation setting is not necessarily straightforward



Automatic Evaluation: Challenges

TASKS

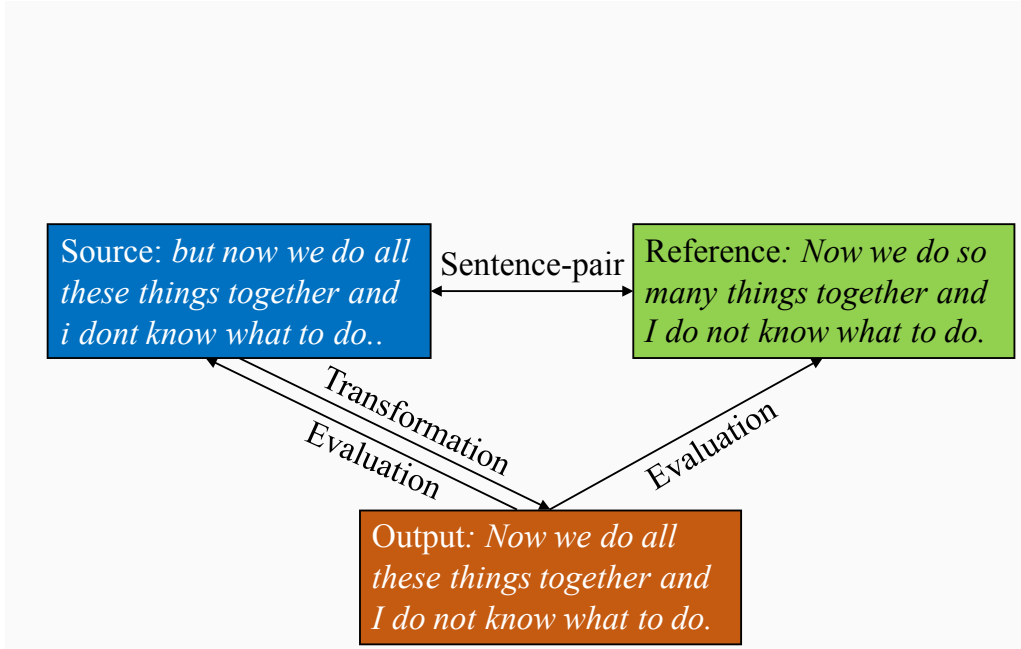
Tasks are conflated under the TST label while they are **not** exactly the same [Lai et al., 2021]

SETTINGS

The evaluation setting is not necessarily straightforward

CONDITIONS

It is unclear how the used metrics correlate to human judgements under different conditions.



Automatic Evaluation: Challenges

TASKS

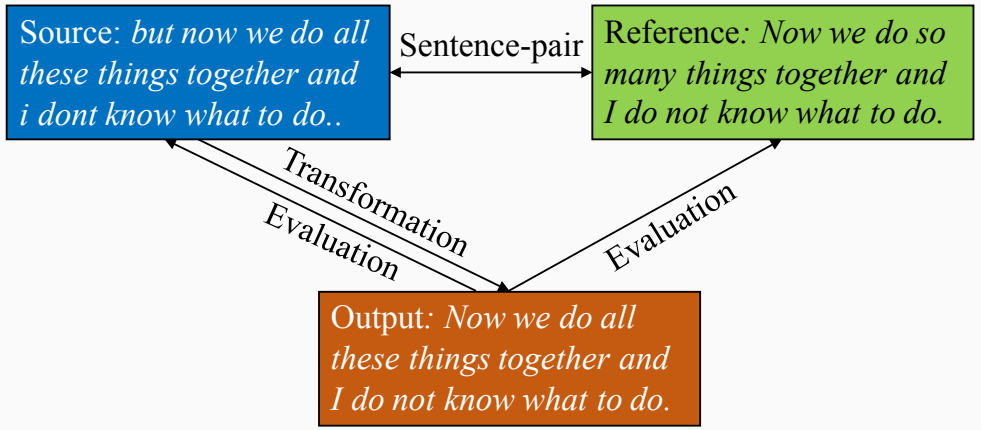
Tasks are conflated under the TST label while they are **not** exactly the same [Lai et al., 2021]

SETTINGS

The evaluation setting is not necessarily straightforward

CONDITIONS

It is unclear how the used metrics correlate to human judgements under different conditions.



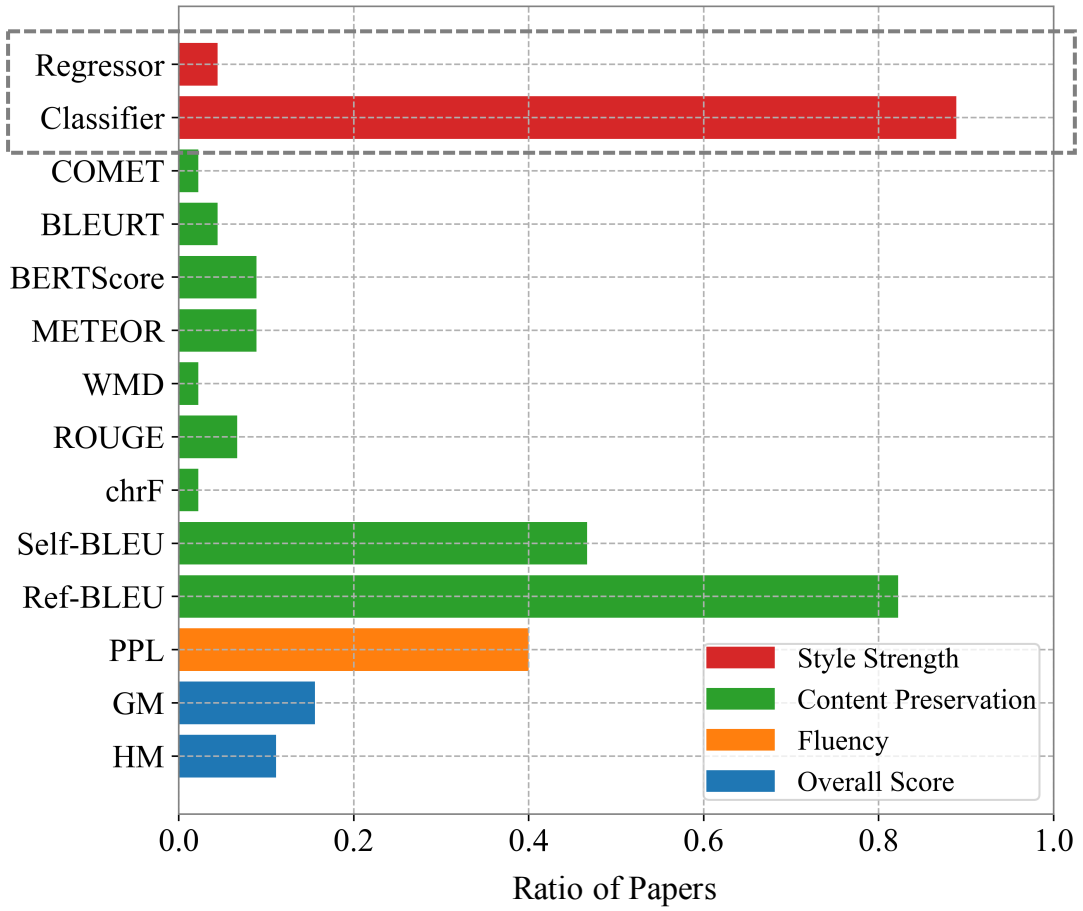
Informal-to-formal

VS

Formal-to-Informal

Formality Transfer: Research Questions

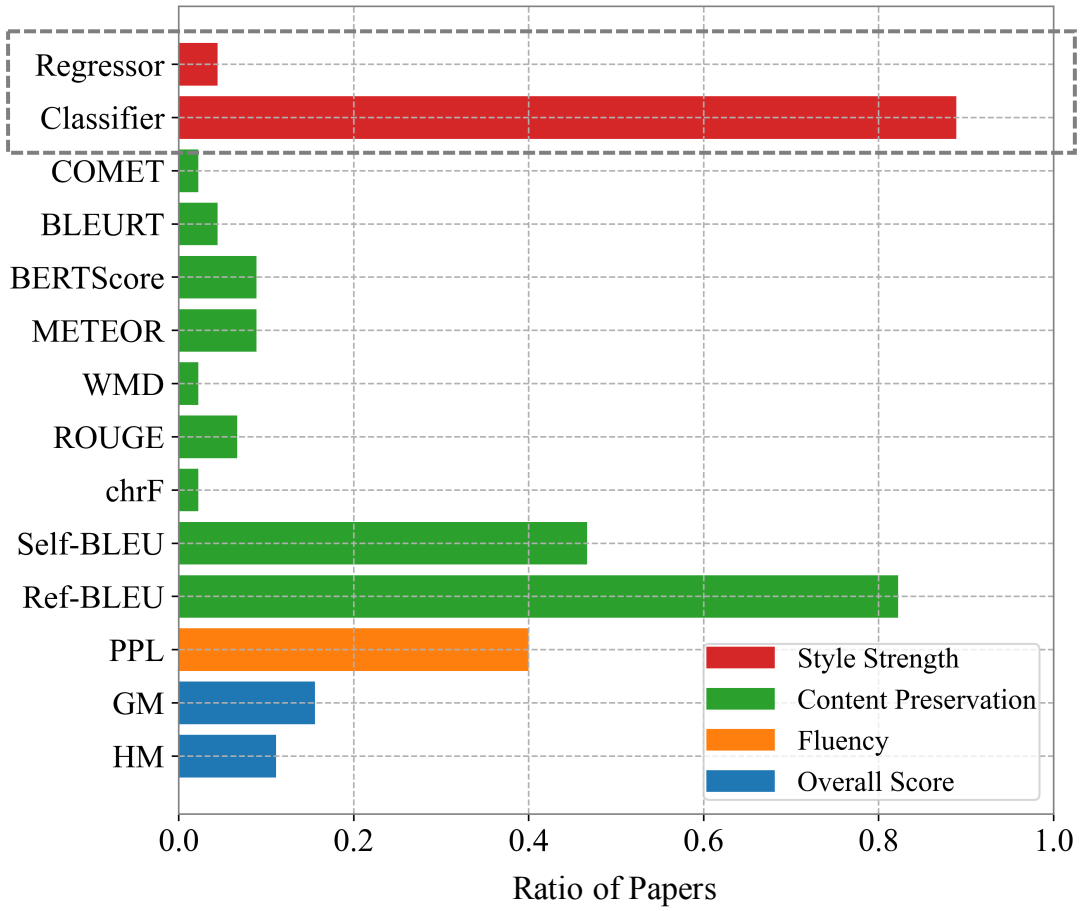
Formality Transfer: Research Questions



CLASSIFIER VS REGRESSOR

How do the classifier and regressor correlate with human judgement?

Formality Transfer: Research Questions



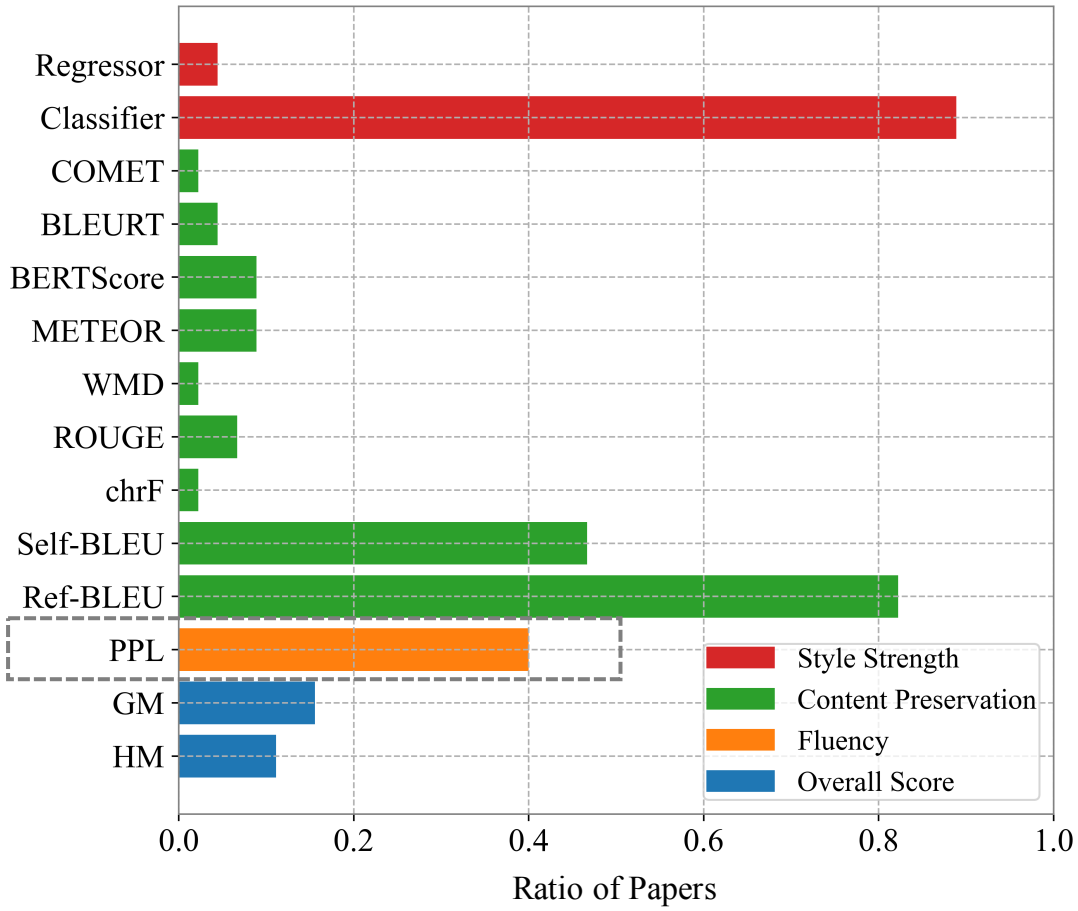
CLASSIFIER VS REGRESSOR

How do the classifier and regressor correlate with human judgement?

SOURCE VS REFERENCE

How do metrics behave when used to compare outputs to source or reference?

Formality Transfer: Research Questions



CLASSIFIER VS REGRESSOR

How do the classifier and regressor correlate with human judgement?

SOURCE VS REFERENCE

How do metrics behave when used to compare outputs to source or reference?

INFORMAL VS FORMAL

Is fluency well captured by perplexity, and what if the target style is informal?

Human Judgement as a Compass

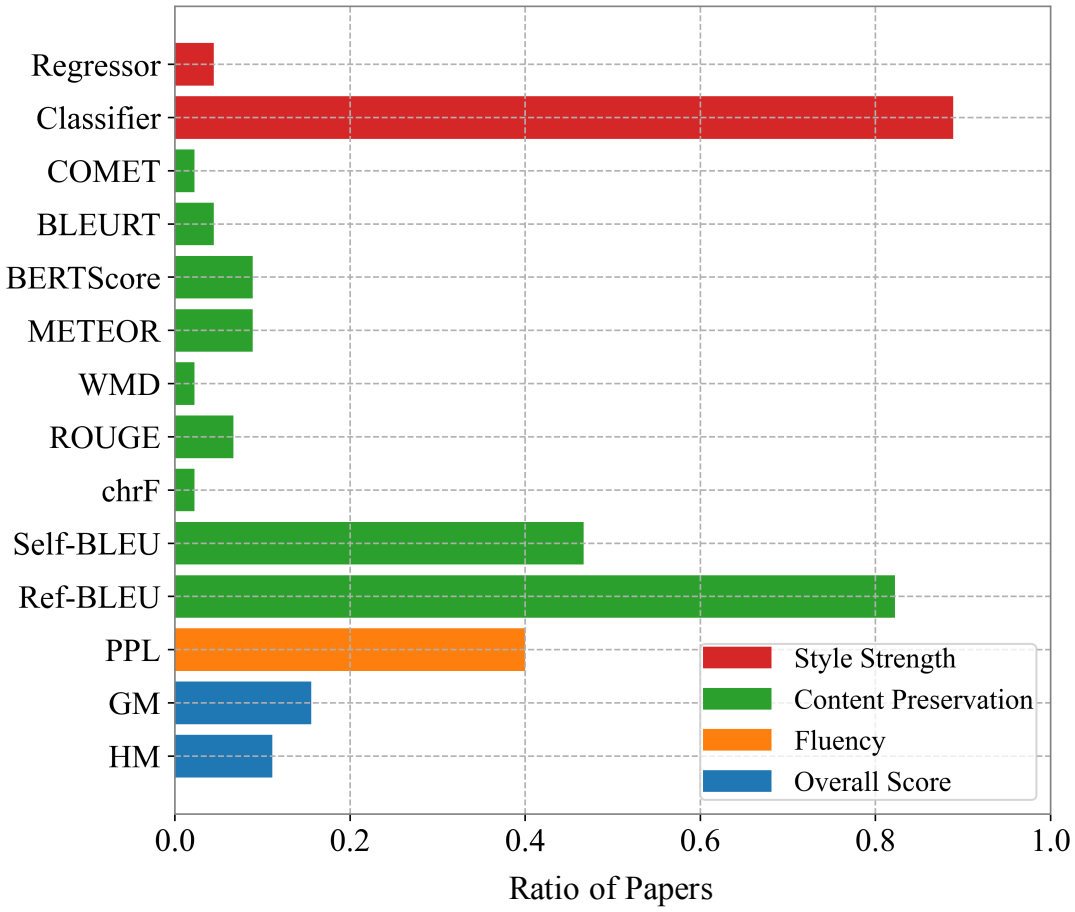
Does the transformed sentence fit the target style?

Is the content of the transformed sentence the same as the original sentence?

Considering the target style, could the transformed sentence have been written by a native speaker?

RAO (Rao and Tetreault, 2018)	Supervised
NIU (Niu et al., 2018)	
BART (Lai et al., 2021b)	
HIGH (Lai et al., 2021a)	
LUO (Luo et al., 2019)	Unsupervised
YI (Yi et al., 2020)	
Zhou (Zhou et al., 2020)	
IBT (Lai et al., 2021a)	
Human Reference	

Human Judgement as a Compass



Correlation Analysis

- Pearson Correlation (system-level)
- Kendall's Tau-like formulation (segment-level)

Human Judgement as a Compass

80 source sentences, **640** output sentences in total.

Task: Rating the transferred sentence on a continuous scale (0-100)

Human Judgement as a Compass

80 source sentences, **640** output sentences in total.

Task: Rating the transferred sentence on a continuous scale (0-100)

university of groningen

Task Guideline

This task consists of judging sentence changes. To this aim, you will be shown different changes in a given sentence. The changes are related to style: from informal to formal or from formal to informal. We call these changes transformations.

There are two examples on this page. For each example, there are two evaluations, which have the same original sentence but different transformed sentences.

Example 1:

In the example below, the sentence is transformed from **informal** to **formal,** and you need to assess if this transformation is indeed correct, i.e. the second sentence is more formal than the first.

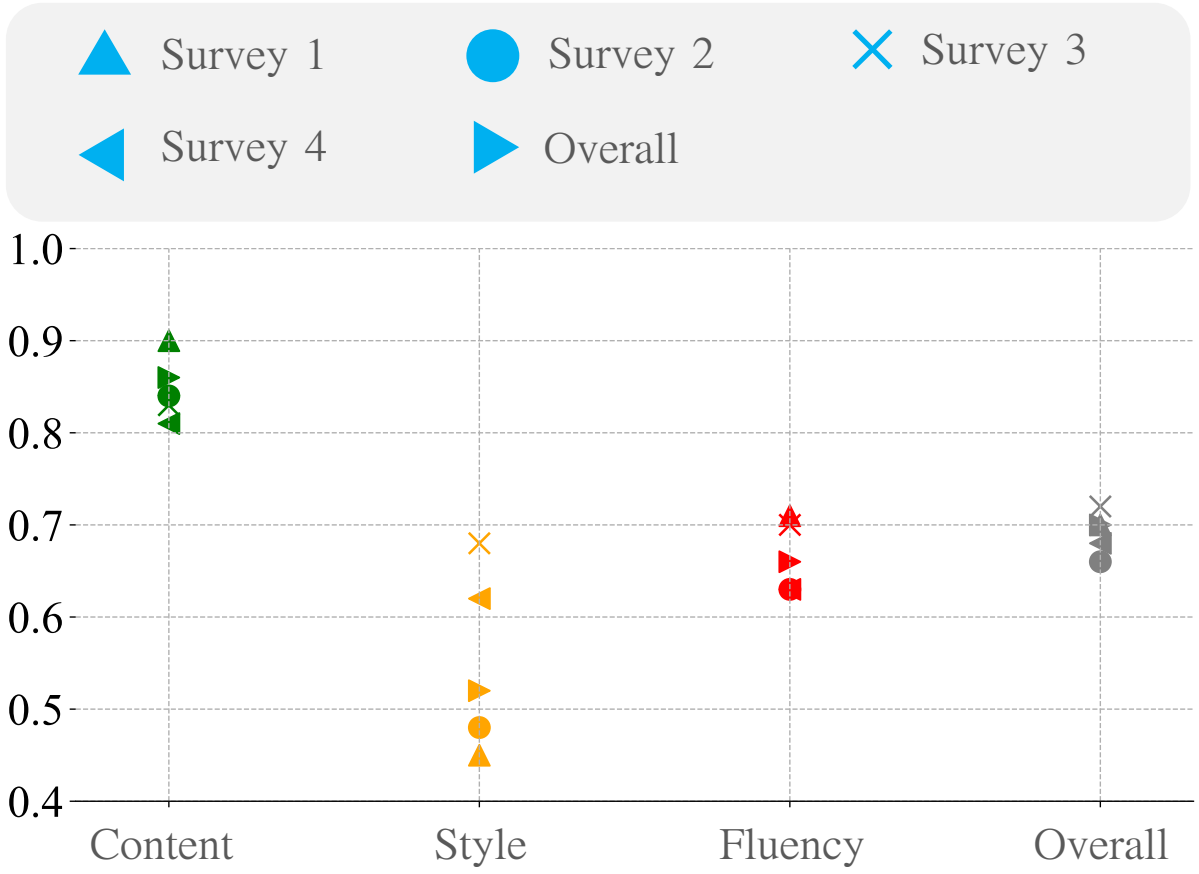
Original sentence:
different from what I've seen though.
Transformed sentence 1:
It differs from what I have seen, however.

The following are plausible evaluations for the transformation:

Strongly disagree — Strongly agree

The content of the rewritten sentence is the same as the given sentence.

Human Judgement as a Compass



Inter-Annotator Agreement

university of groningen

Task Guideline

This task consists of judging sentence changes. To this aim, you will be shown different changes in a given sentence. The changes are related to style: from informal to formal or from formal to informal. We call these changes transformations.

There are two examples on this page. For each example, there are two evaluations, which have the same original sentence but different transformed sentences.

Example 1:

In the example below, the sentence is transformed from **informal** to **formal,** and you need to assess if this transformation is indeed correct, i.e. the second sentence is more formal than the first.

Original sentence:
different from what I've seen though.
Transformed sentence 1:
It differs from what I have seen, however.

The following are plausible evaluations for the transformation:

Strongly disagree — Strongly agree

The content of the rewritten sentence is the same as the given sentence.

Automatic Evaluation: Style Strength

- **Based Model**
  Fine-tuning pre-trained model BERT
- **Training Data**
  Rating data of PT16 [Pavlick et al., 2016]
  Style labelled data of GYAFC [Rao et al., 2018] or PT16
- **C-GYAFC: 94.4%; C-PT16: 58.6%**

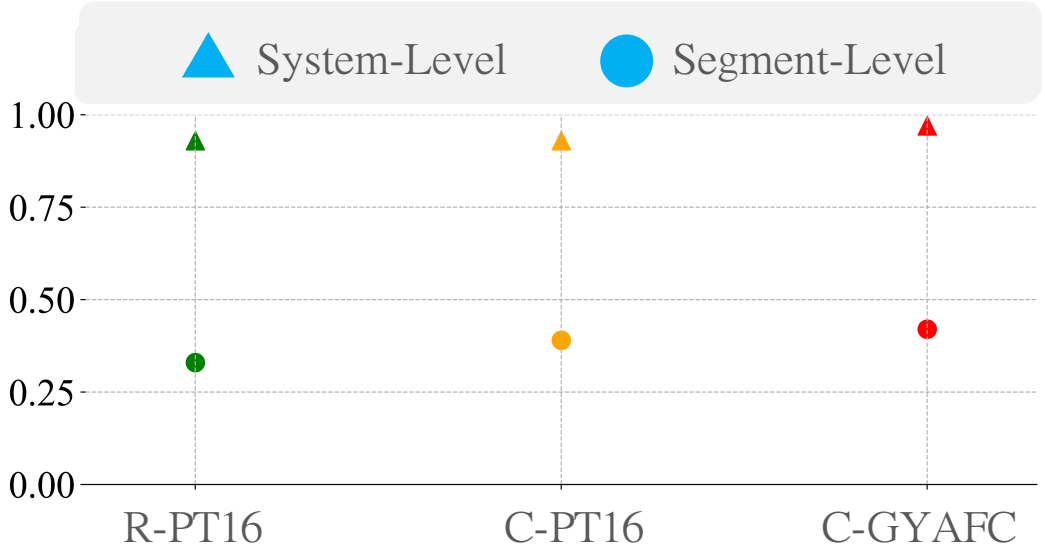
Automatic Evaluation: Style Strength

- **Based Model**
  Fine-tuning pre-trained model BERT
- **Training Data**
  Rating data of PT16 [Pavlick et al., 2016]
  Style labelled data of GYAFC [Rao et al., 2018] or PT16
- **Regressor (R) VS Classifier (C)**

RQ: How do the classifier and regressor correlate with human judgement?

Automatic Evaluation: Style Strength

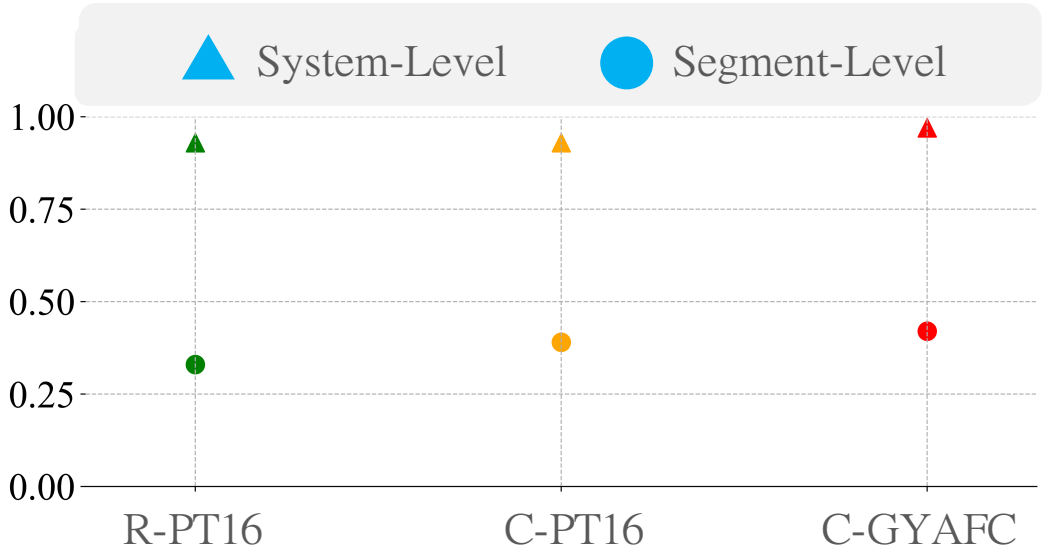
- **Based Model**
  Fine-tuning pre-trained model BERT
- **Training Data**
  Rating data of PT16 [Pavlick et al., 2016]
  Style labelled data of GYAFC [Rao et al., 2018] or PT16
- **Regressor (R) VS Classifier (C)**

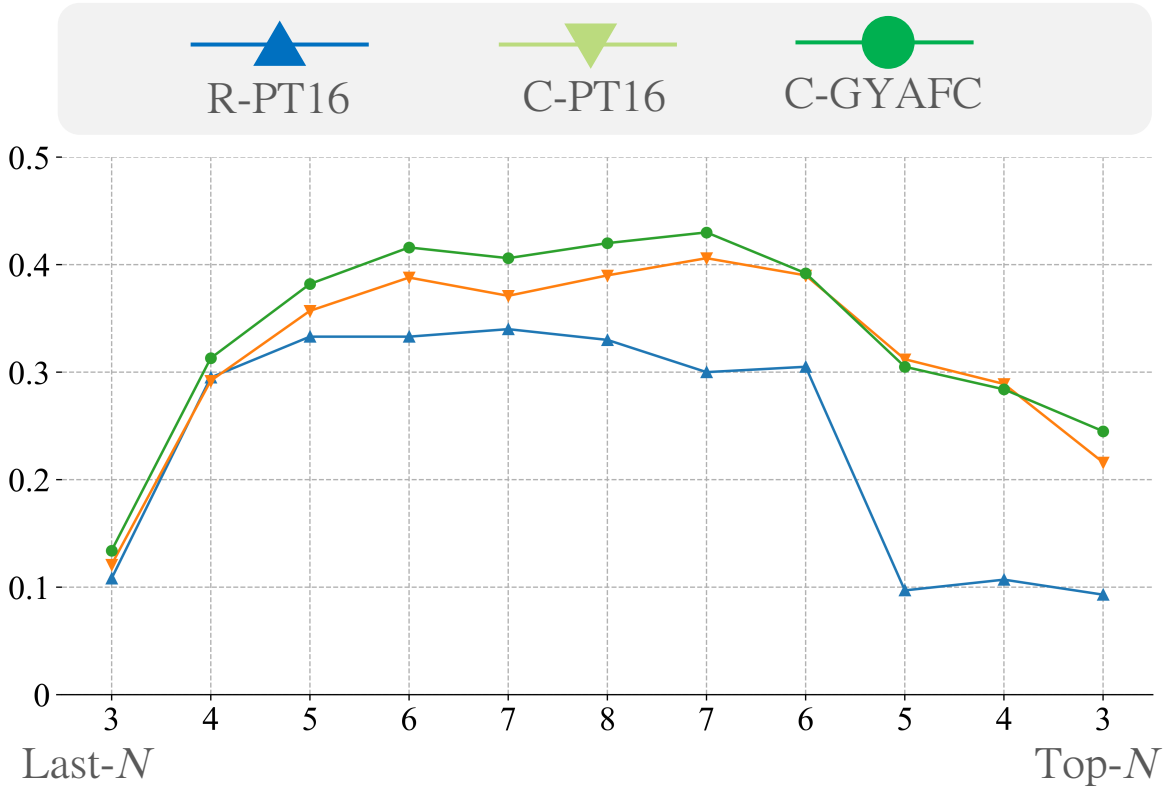


Correlation of automatic metrics in style strength with human judgements.

Automatic Evaluation: Style Strength

- **Based Model**
  Fine-tuning pre-trained model BERT
- **Training Data**
  Rating data of PT16
  Style labelled data of GYAFC or PT16
- **Regressor (R) VS Classifier (C)**





Correlation computed over the top-/last-*N* systems which are sorted by human judgements.

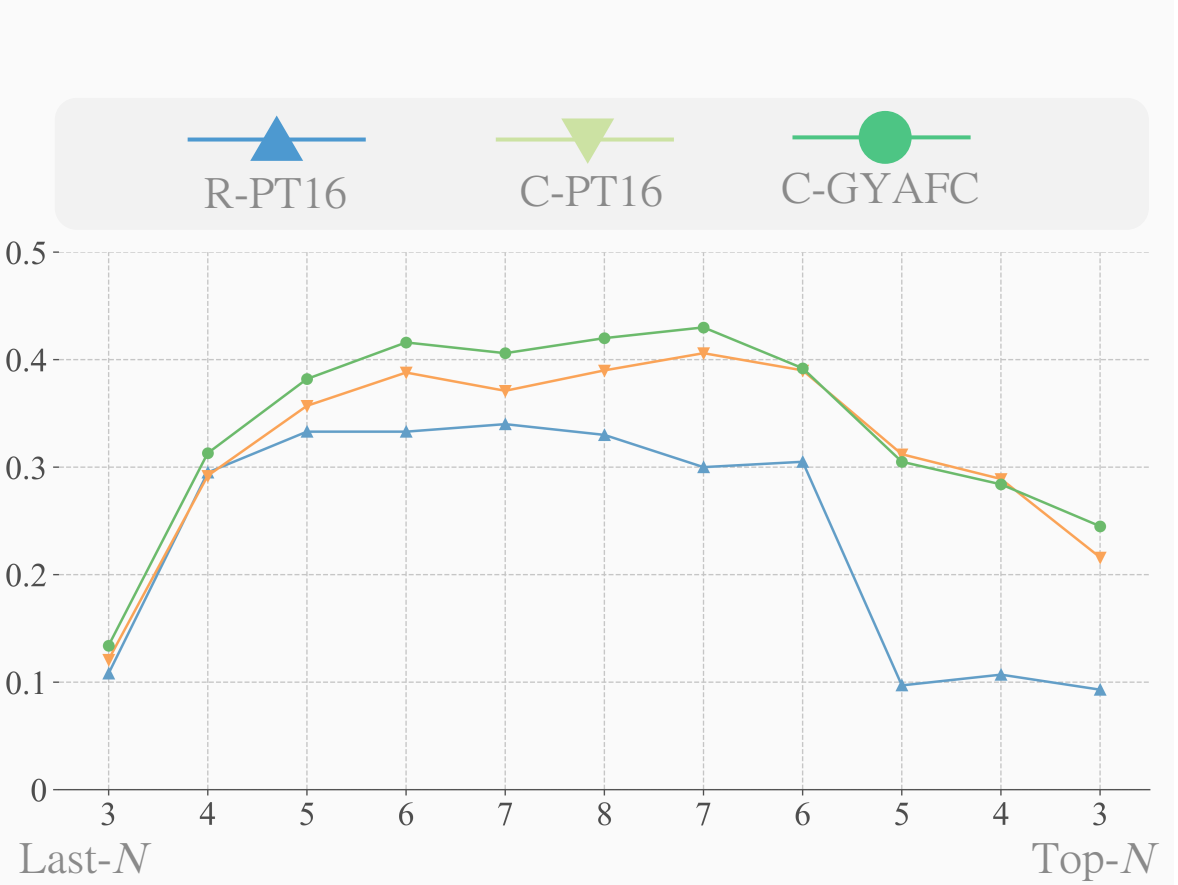
Automatic Evaluation: Style Strength

- **Based Model**
  Fine-tuning pre-trained model BERT
- **Training Data**
  Rating data of PT16
  Style labelled data of GYAFC or PT16
- **Regressor (R) VS Classifier (C)**

Style Strength

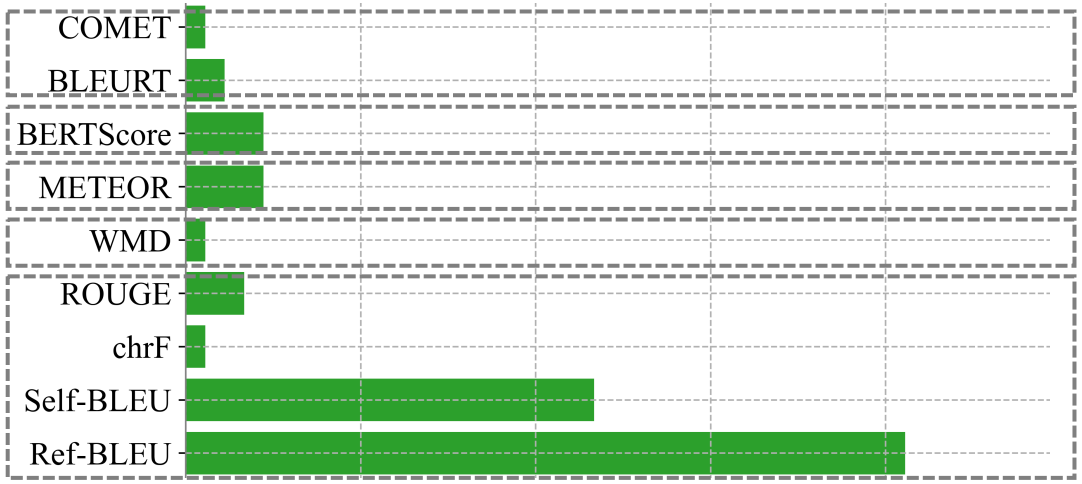
× Style regressor performs worse when evaluating high-quality TST systems

✓We recommend the classifier with the highest performance.



Correlation computed over the top-/last-*N* systems which are sorted by human judgements.

Automatic Evaluation: Content Preservation



Learnable

Neural based

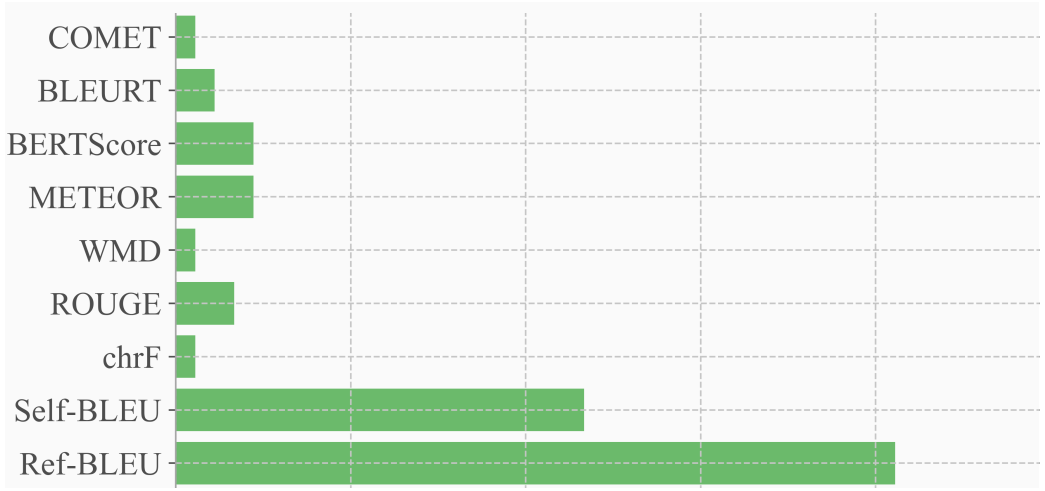
String based

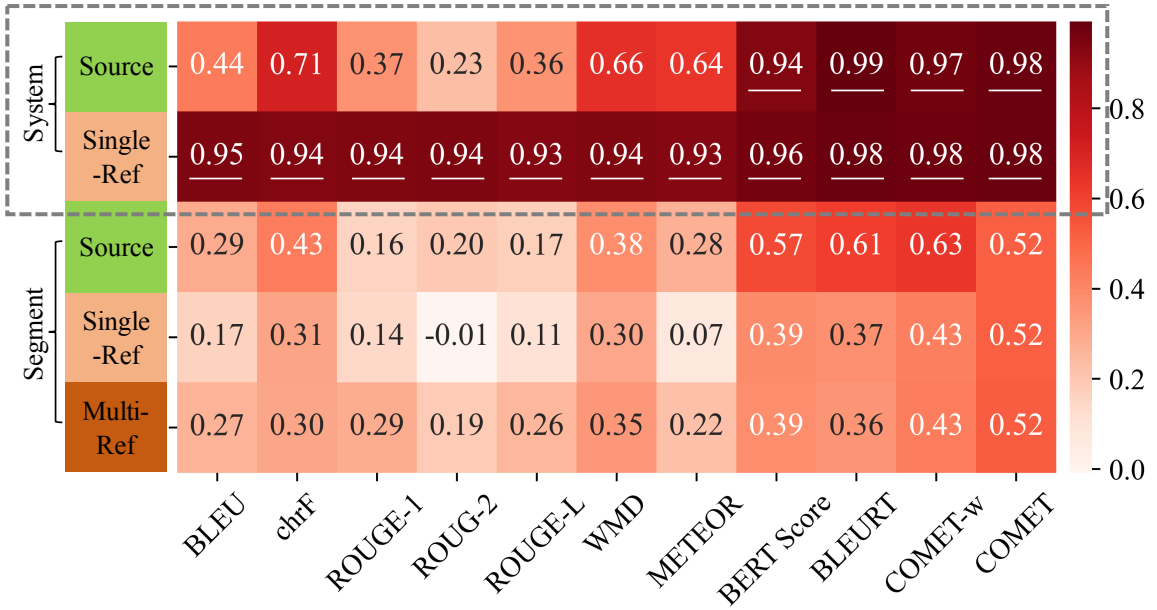
Embedding based

N-gram based

RQ: How do automatic metrics behave when used to compare outputs to source or reference?

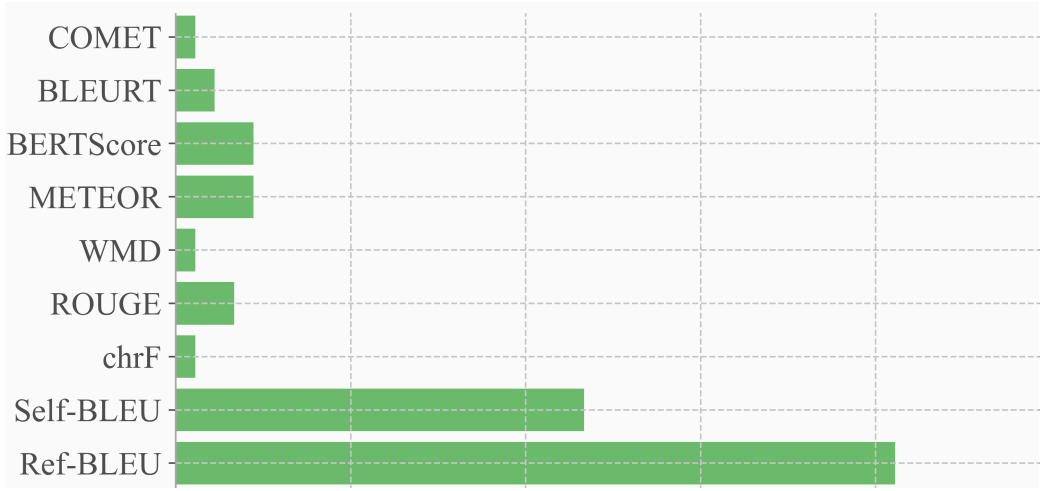
Automatic Evaluation: Content Preservation

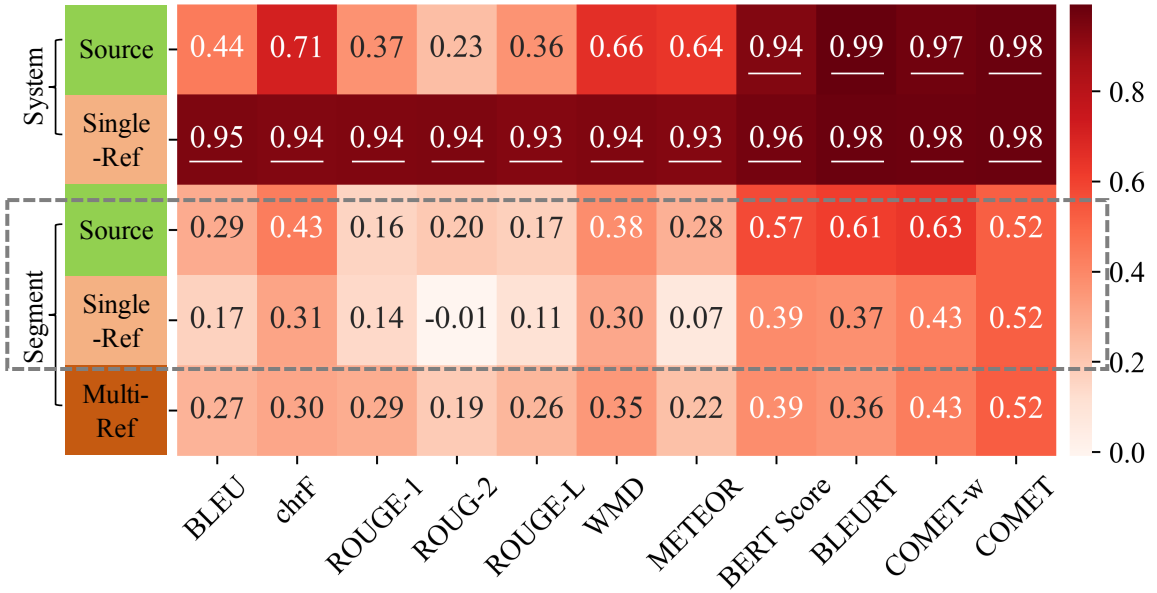




Correlations of automatic metrics computed against source/reference with human judgments. Underlining indicates $p < 0.01$.

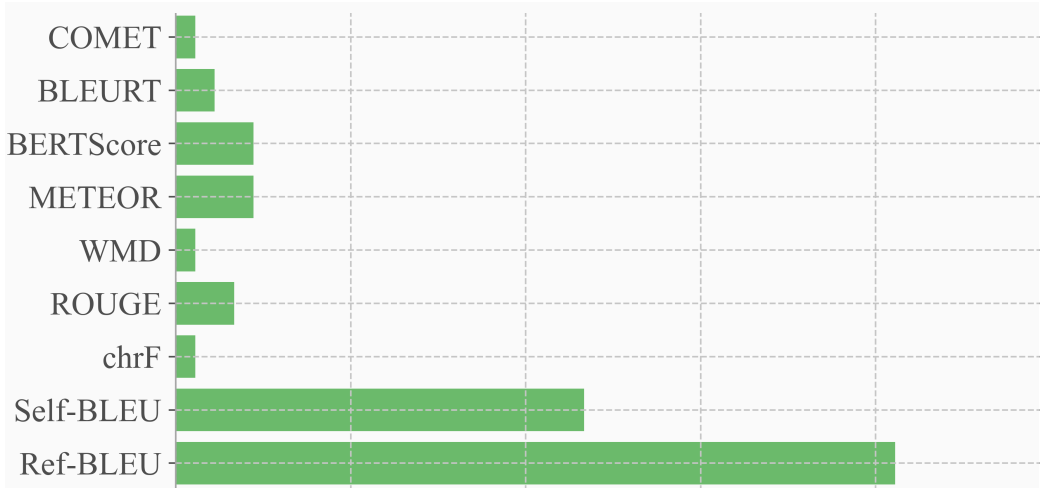
Automatic Evaluation: Content Preservation

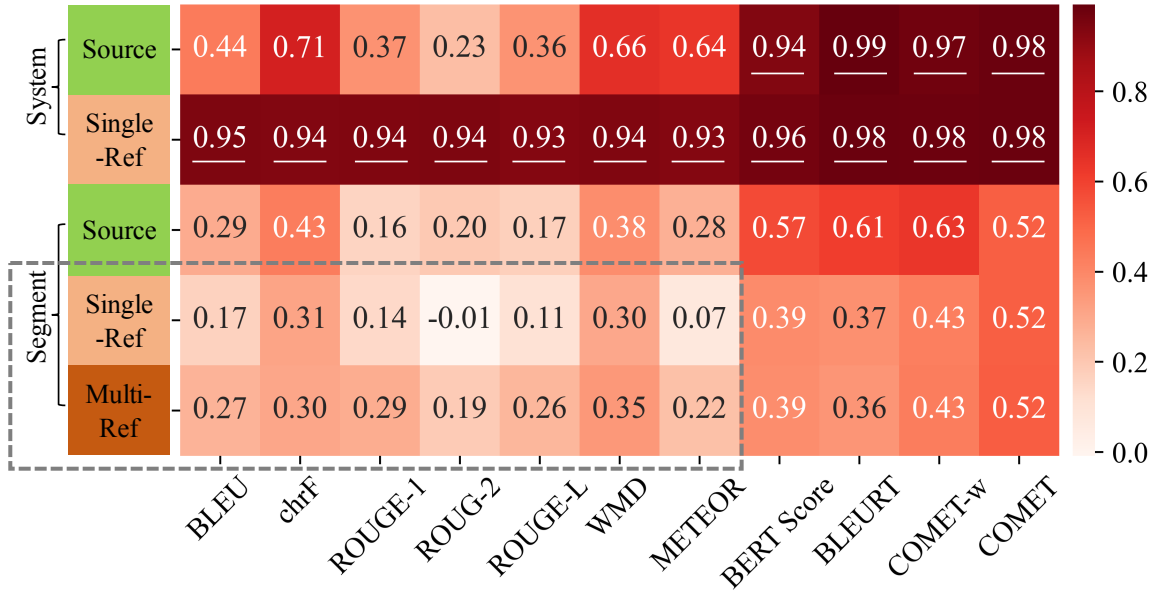




Correlations of automatic metrics computed against source/reference with human judgments. Underlining indicates $p < 0.01$.

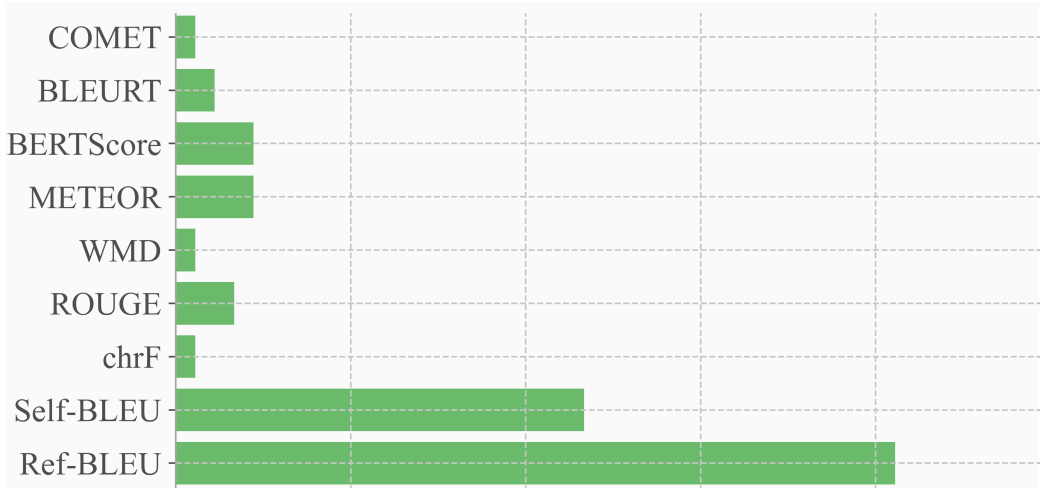
Automatic Evaluation: Content Preservation





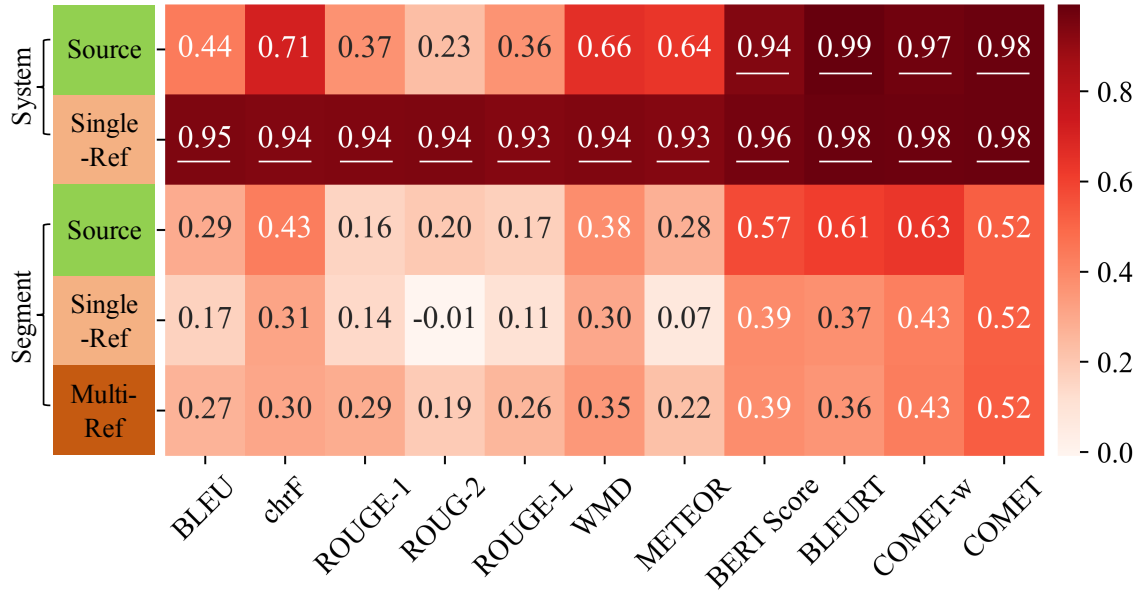
Correlations of automatic metrics computed against source/reference with human judgments. Underlining indicates $p < 0.01$.

Automatic Evaluation: Content Preservation



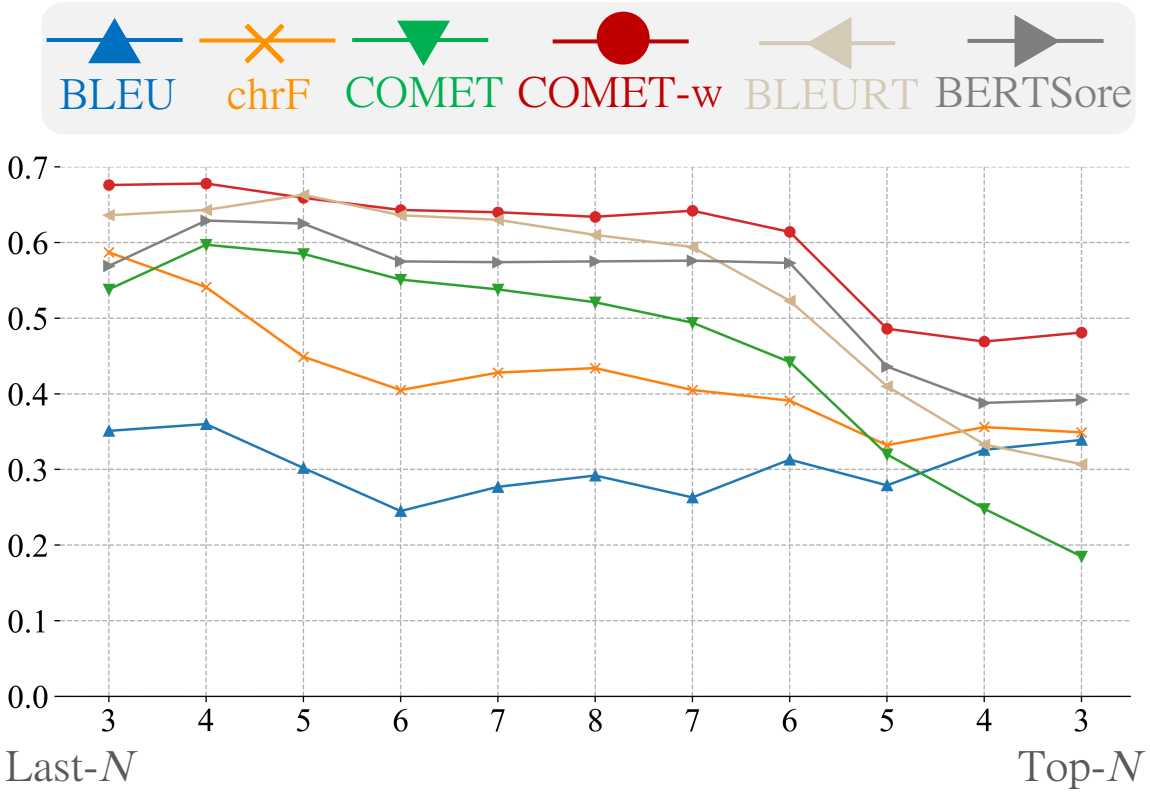
	BLEU	chrF	ROUGE-1	ROUGE-2	ROUGE-L	WMD	METEOR	BERTScore	BLEURT	COMET-w
Reference 2	0.28	0.37	0.33	0.10	0.36	0.46	0.21	0.59	0.61	0.61
Reference 3	0.25	0.41	0.37	0.12	0.35	0.47	0.34	0.60	0.60	0.55
Reference 4	0.37	0.41	0.46	0.24	0.46	0.49	0.31	0.60	0.56	0.62

Correlation between using the first human reference and others at segment-level.

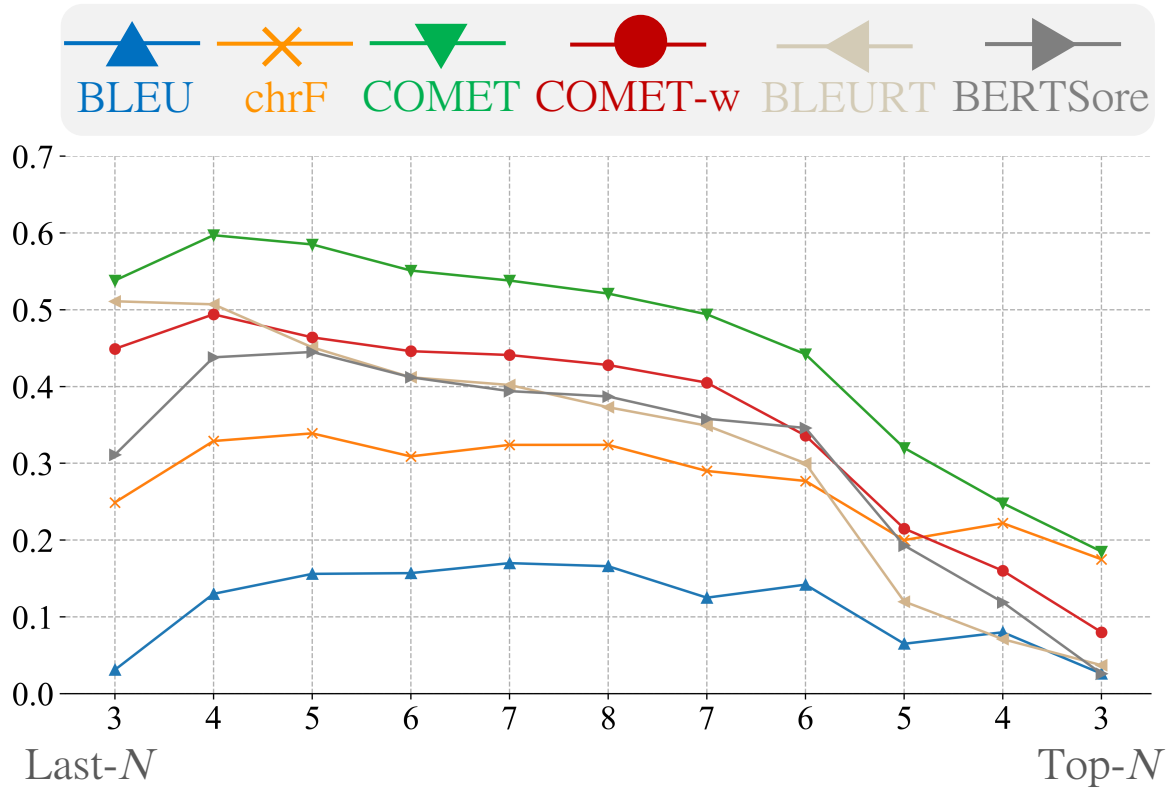


Correlations of automatic metrics computed against source/reference with human judgments. Underlining indicates $p < 0.01$.

Automatic Evaluation: Content Preservation



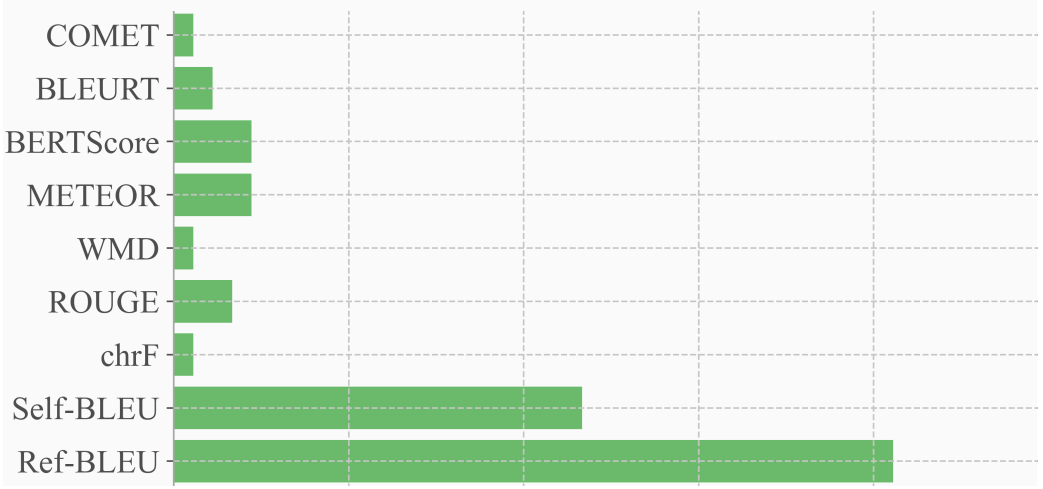
Automatic results against source sentence.



Automatic results against human reference.

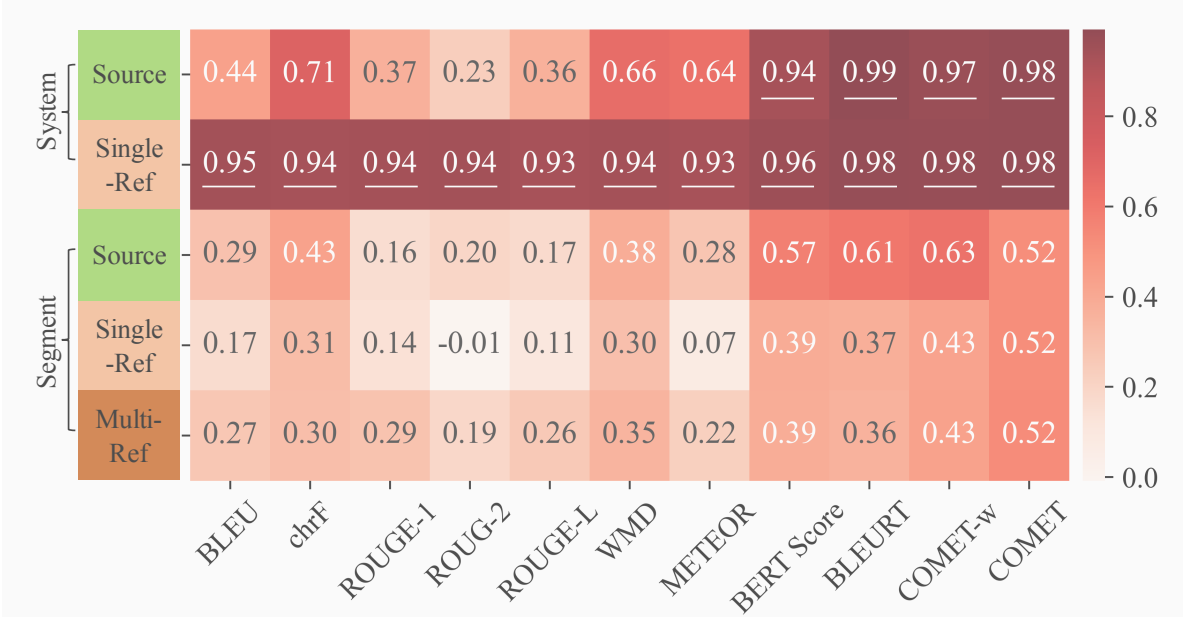
Correlation in content computed over the top-/last-*N* systems sorted by human judgements.

Evaluation Practices: Content Preservation



Content Preservation

- × Segment-level — surface-base metrics
- ✓ Using the source — Learnable metrics
- ✓ Using the reference at system-level— Most metrics are reliable



Correlations of automatic metrics computed against source/reference with human judgments. Underlining indicates $p < 0.01$.

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

	N	Informal-to-Formal	Formal-to-Informal
System-level (r)	8	<u>0.96</u>	0.65
Segment-level (τ)	320	0.52	0.35

Absolute correlation of automatic metrics with human judgements. The underlined scores indicate $p < 0.01$.

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

	N	Informal-to-Formal	Formal-to-Informal
System-level (r)	8	0.96	0.65
Segment-level (τ)	320	0.52	0.35

Absolute correlation of automatic metrics with human judgements. The underlined scores indicate $p < 0.01$.

	Informal-to-Formal			Formal-to-Informal		
	GPT2-Inf	GPT2-For	r	GPT2-Inf	GPT2-For	r
Source	76	143	-	87	68	-
Reference	60	37	0.21	115	270	0.13
BART	34	**26**	0.33	**24**	**28**	0.02
IBT	**32**	**26**	0.32	33	40	0.17
NIU	43	37	0.30	71	75	0.03
HIGH	41	35	**0.62**	80	75	0.00
RAO	54	57	0.33	54	55	0.02
ZHOU	189	218	0.36	103	111	**0.42**
YI	160	182	0.31	205	436	0.27
LUO	128	152	0.43	6962	8191	0.17

GPT-2 based perplexity scores and their Pearson correlation with human judgements at segment-level.

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

	N	Informal-to-Formal	Formal-to-Informal
System-level (r)	8	0.96	0.65
Segment-level (τ)	320	0.52	0.35

Absolute correlation of automatic metrics with human judgements. The underlined scores indicate $p < 0.01$.

	Informal-to-Formal			Formal-to-Informal		
	GPT2-Inf	GPT2-For	r	GPT2-Inf	GPT2-For	r
Source	76	143	-	87	68	-
Reference	60	37	0.21	115	270	0.13
BART	34	**26**	0.33	**24**	**28**	0.02
IBT	**32**	**26**	0.32	33	40	0.17
NIU	43	37	0.30	71	75	0.03
HIGH	41	35	**0.62**	80	75	0.00
RAO	54	57	0.33	54	55	0.02
ZHOU	189	218	0.36	103	111	**0.42**
YI	160	182	0.31	205	436	0.27
LUO	128	152	0.43	6962	8191	0.17

GPT-2 based perplexity scores and their Pearson correlation with human judgements at segment-level.

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

	N	Informal-to-Formal	Formal-to-Informal
System-level (r)	8	0.96	0.65
Segment-level (τ)	320	0.52	0.35

Absolute correlation of automatic metrics with human judgements. The underlined scores indicate $p < 0.01$.

	Informal-to-Formal			Formal-to-Informal		
	GPT2-Inf	GPT2-For	r	GPT2-Inf	GPT2-For	r
Source	76	143	-	87	68	-
Reference	60	37	0.21	115	270	0.13
BART	34	**26**	0.33	**24**	**28**	0.02
IBT	**32**	**26**	0.32	33	40	0.17
NIU	43	37	0.30	71	75	0.03
HIGH	41	35	**0.62**	80	75	0.00
RAO	54	57	0.33	54	55	0.02
ZHOU	189	218	0.36	103	111	**0.42**
YI	160	182	0.31	205	436	0.27
LUO	128	152	0.43	6962	8191	0.17

GPT-2 based perplexity scores and their Pearson correlation with human judgements at segment-level.

Automatic Evaluation: Fluency

- **Based Model**
  Fine-tuning pre-trained model GPT-2
- **Training Data**
  Formal texts—GPT2-For (informal-to-formal)
  Informal texts—GPT2-Inf (formal-to-informal)
- **Metric**
  Perplexity

	Informal-to-Formal			Formal-to-Informal		
	GPT2-Inf	GPT2-For	r	GPT2-Inf	GPT2-For	r
Source	76	143	-	87	68	-
Reference	60	37	0.21	115	270	0.13
BART	34	**26**	0.33	**24**	**28**	0.02
IBT	**32**	**26**	0.32	33	40	0.17
NIU	43	37	0.30	71	75	0.03
HIGH	41	35	**0.62**	80	75	0.00
RAO	54	57	0.33	54	55	0.02
ZHOU	189	218	0.36	103	111	**0.42**
YI	160	182	0.31	205	436	0.27
LUO	128	152	0.43	6962	8191	0.17

GPT-2 based perplexity scores and their Pearson correlation with human judgements at segment-level.

Fluency

× Perplexity is clearly less reliable for the formal-to-informal direction

✓ Perplexity can be used for evaluating the informal-to-formal direction

Broader Implications for Style Transfer

Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Polarity Swap

Target: *i hate this screen, it is not the right size*

Broader Implications for Style Transfer

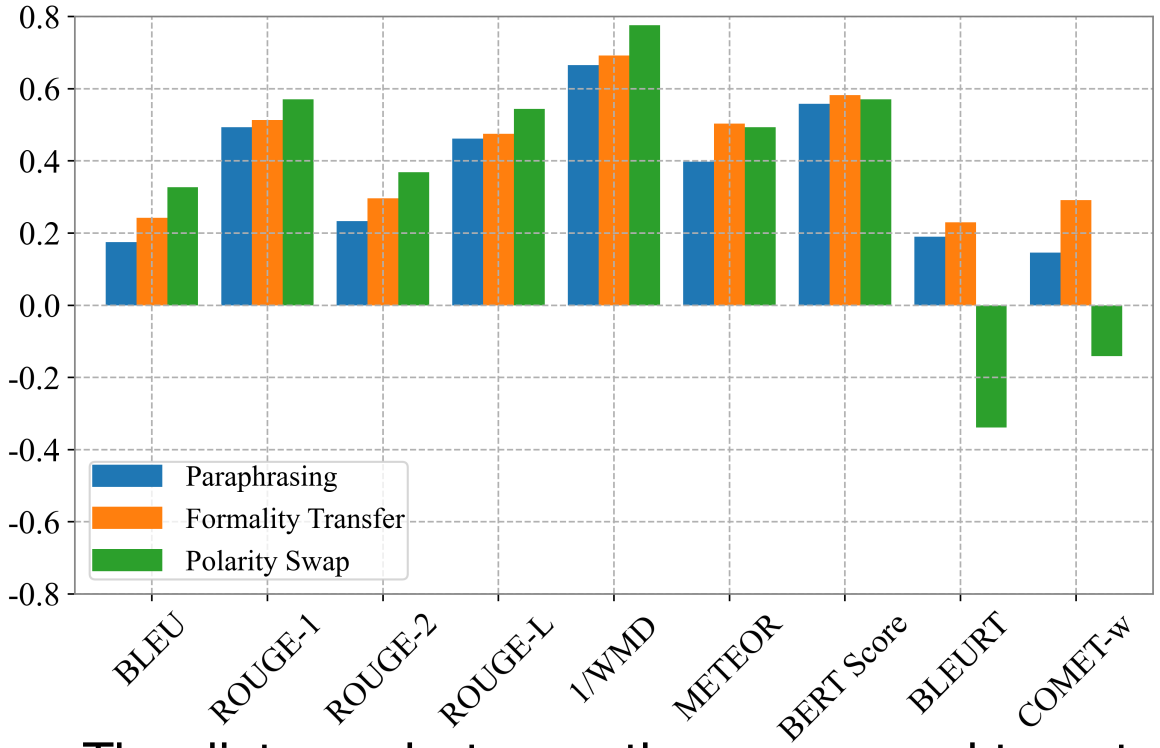
Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Polarity Swap

Target: *i hate this screen, it is not the right size*



The distance between the source and target sentences measured by content metrics.

Broader Implications for Style Transfer

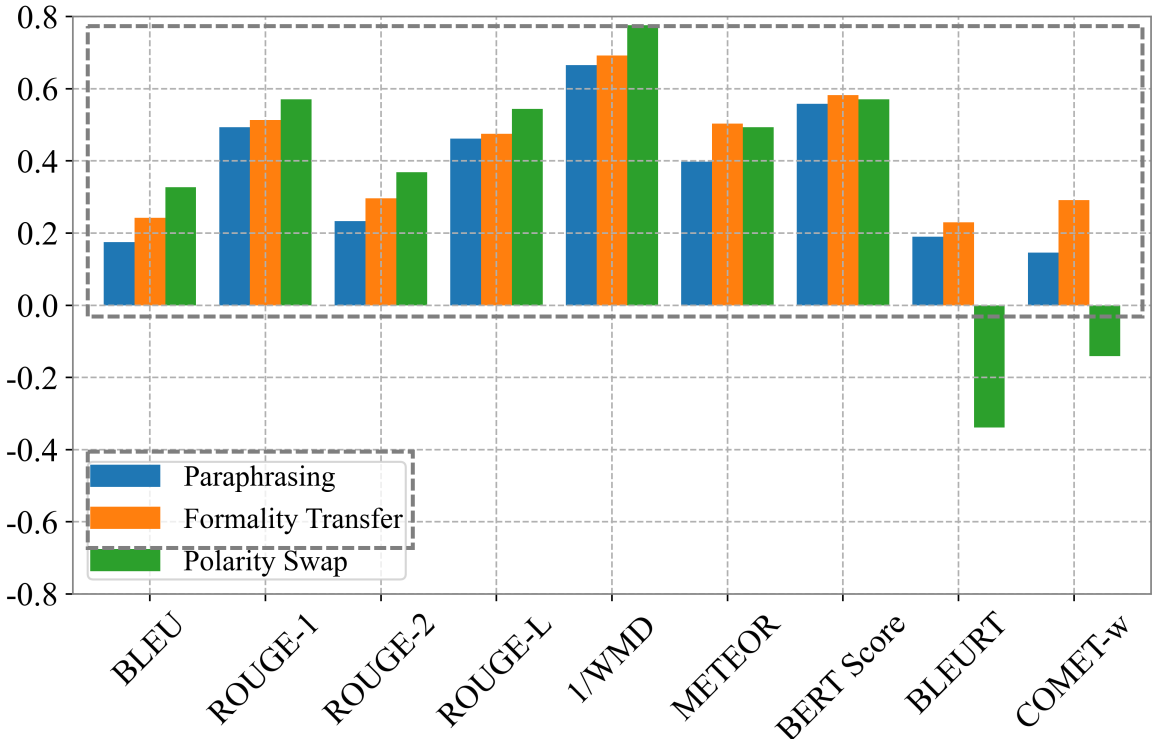
Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Polarity Swap

Target: *i hate this screen, it is not the right size*



The distance between the source and target sentences measured by content metrics.

Broader Implications for Style Transfer

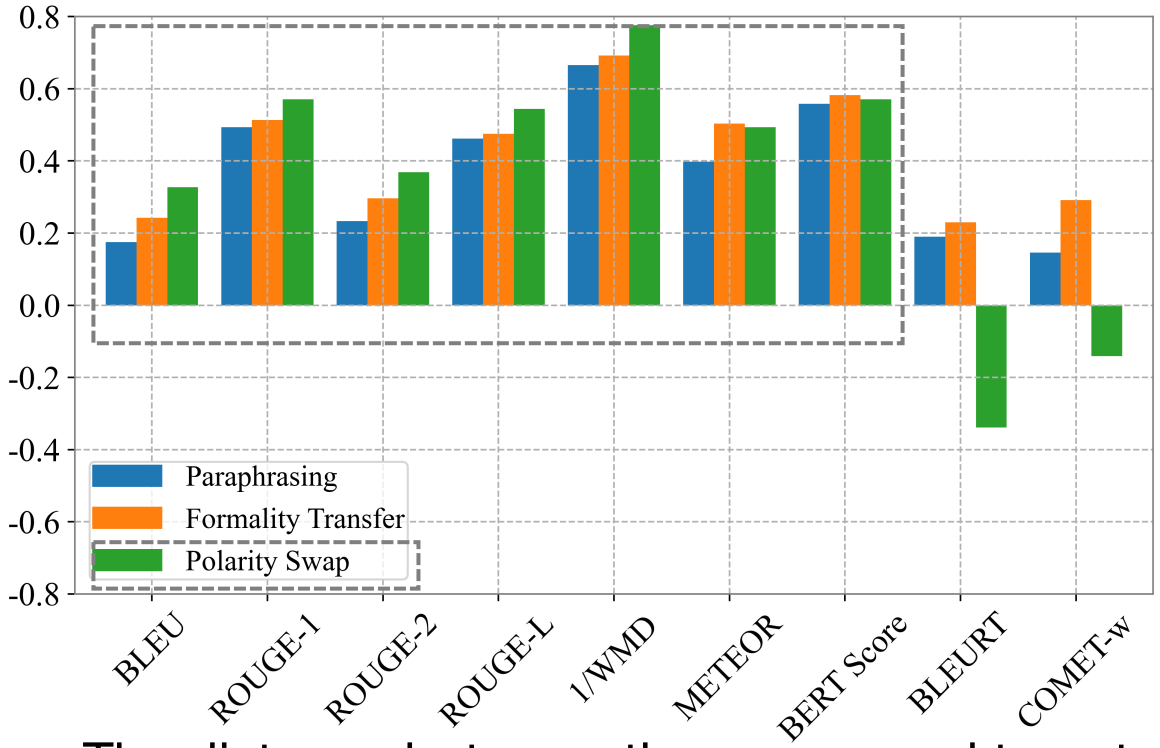
Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Polarity Swap

Target: *i hate this screen, it is not the right size*



The distance between the source and target sentences measured by content metrics.

Broader Implications for Style Transfer

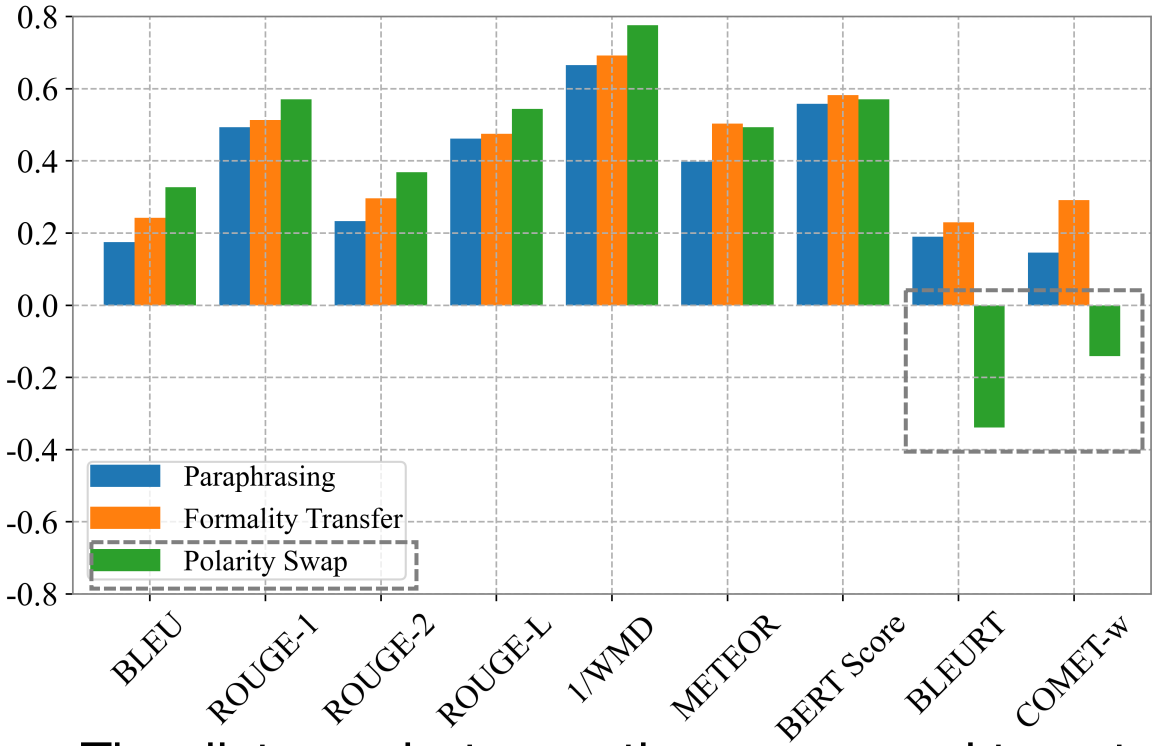
Source: *i like this screen, it's just the right size...*

Formality Transfer

Target: *I like this screen, it is just the right size.*

Polarity Swap

Target: *i hate this screen, it is not the right size*

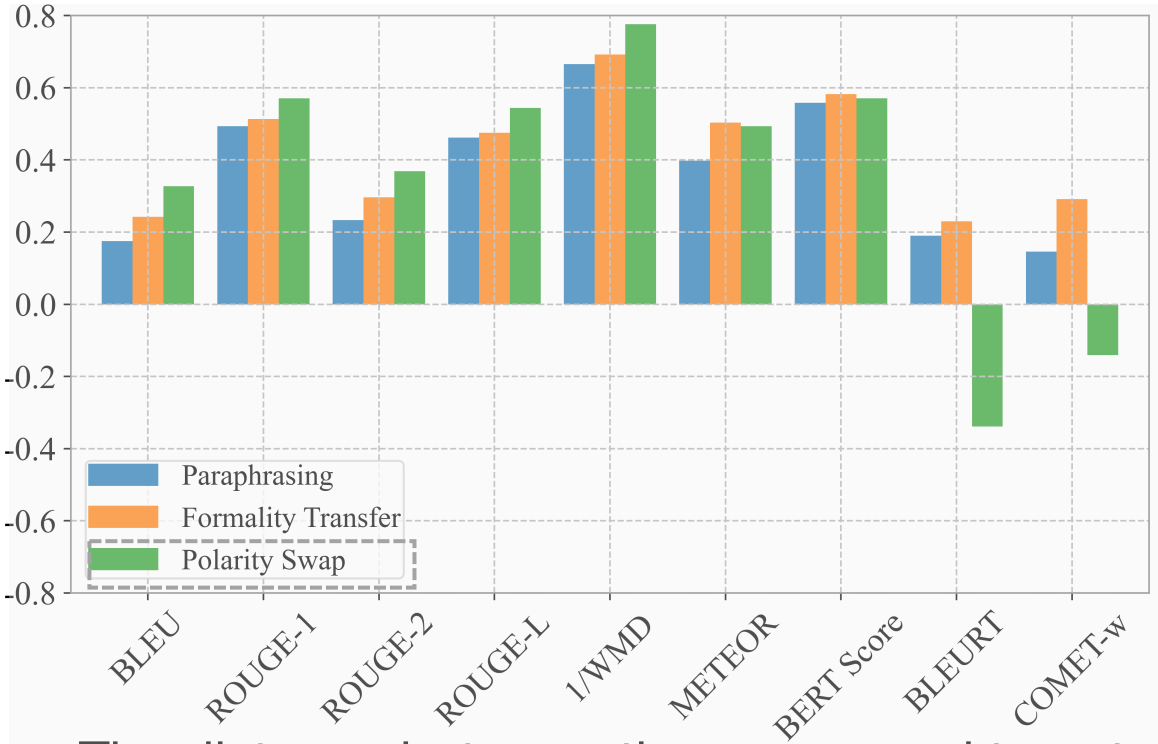


The distance between the source and target sentences measured by content metrics.

Broader Implications for Style Transfer

Polarity Swap

- How to best use these metrics in polarity swap under different settings (e.g. using source vs reference)



The distance between the source and target sentences measured by content metrics.

Take-Home Message!

Formality Transfer:

- ✓ **Style Strength:** Classifier with the highest performance
- ✓ Content Preservation: Learnable Metrics (e.g. COMET)
- ✓ Fluency: Perplexity for informal-to-formal transformation

Contact

@HuiyuanLai @_atoral @MalvinaNissim @GroNlp

huiyuanlai.l@gmail.com

https://arxiv.org/abs/2204.07549

Code, data, and paper list

github.com/laihuiyuan/eval-formality-transfer