



HACETTEPE  
UNIVERSITY  
COMPUTER  
VISION LAB



/mertkilickaya\_



kilickayamert@gmail.com

# Re-evaluating automatic metrics for image captioning

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, Erkut Erdem



**source**

**target**

$$f \left( \text{source image} \right) = \text{The cat sat on the mat}$$

?

# Evaluation

## Human



## Automatic

- Borrowed from Machine Translation
  - BLEU
  - METEOR
  - ROUGE
- Developed for Image Captioning
  - CIDEr
  - SPICE

## Reference

{The cat sat on  
the mat}




## Candidate

{An orange cat  
sitting on mat}

$$BLEU \approx \frac{\{cat, on, mat\}}{\{the, cat, sat, on, the, mat\}}$$

# Existing metrics rely on the overlap between reference and candidate captions


Reference: {The cat sat on the mat}



Candidate: {An orange cat sitting on mat}

$$BLEU \cong \frac{\{cat, on, mat\}}{\{the, cat, sat, on, the, mat\}}$$


Reference: {The cat sat on the mat}



Candidate: {An orange cat sitting on mat}

$$ROUGE \cong \frac{\{cat, on, mat\}}{\{an, orange, cat, sitting, on, mat\}}$$


Reference: {The cat sat on the mat}



Candidate: {An orange kitty sitting on mat}

$$METEOR \cong \frac{\frac{\{kitty, sit, on, mat\}}{\{the, cat, sat, on, the, mat\}} \times \{kitty, sit, on, mat\}}{\{an, orange, cat, sitting, on, mat\}}$$

Reference: {The cat sat on the mat}




Candidate: {An orange kitty sitting on mat}

*CIDEr*  $f(\text{corpus}) =$

Word	Frequency
Kitty	0.6
Sit	0.2
On	0.1
Mat	0.1

Reference: {The cat sat on the mat}



Candidate: {An orange kitty sitting on mat}

$$SPICE \cong f_{\text{objects}} * f_{\text{attributes}} * f_{\text{relations}}$$

What if the captions are not overlapping but still semantically relevant?



Reference

A man wearing a lifevest is sitting in a canoe

Candidate 1

A guy with a red jacket is standing on a boat

Candidate 2

A small white ferry rides through water

We propose: Word Mover Distance

## *Word Mover Distance (WMD)*



Candidate 1

A **guy** with a **red jacket** is **standing** on a **boat**

2.49

= 0.48

+ 0.50

+ 0.60

+ 0.43

+ 0.48

Reference

A **man** **wearing** a **lifevest** is **sitting** in a **canoe**

3.07

=

0.61

+ 0.57

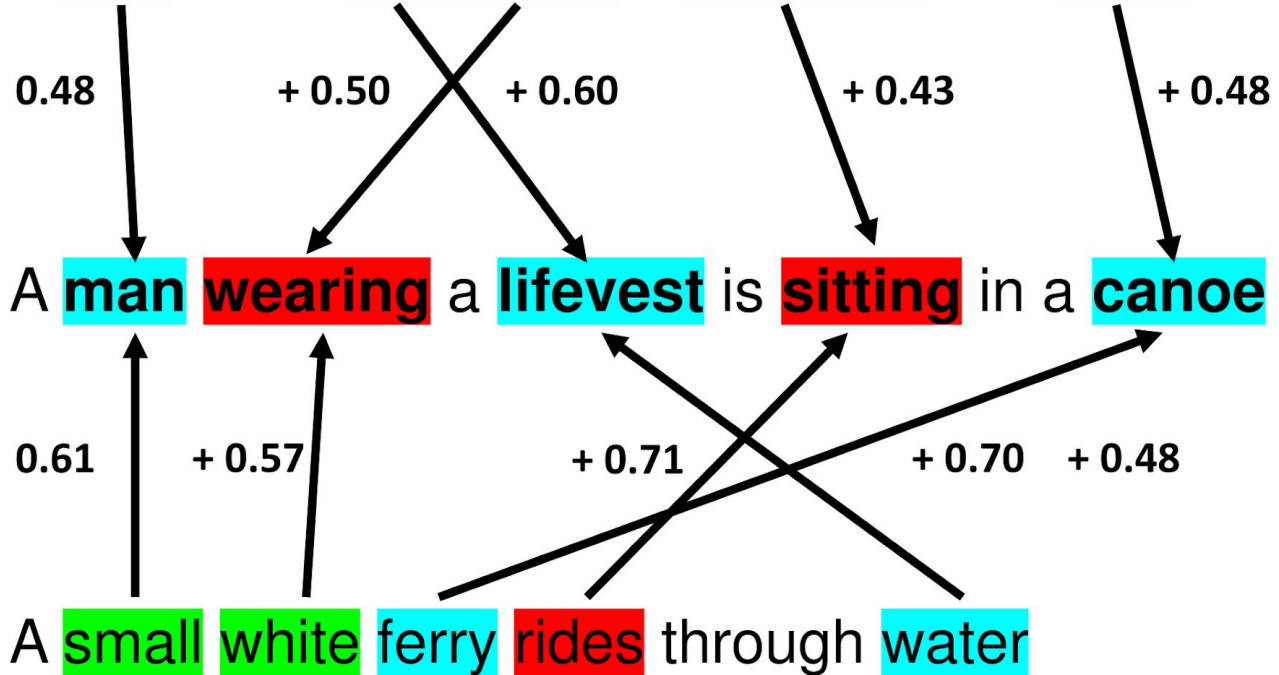
+ 0.71

+ 0.70

+ 0.48

Candidate 2

A **small white** **ferry** **rides** through **water**





So many metrics: Which one is better?

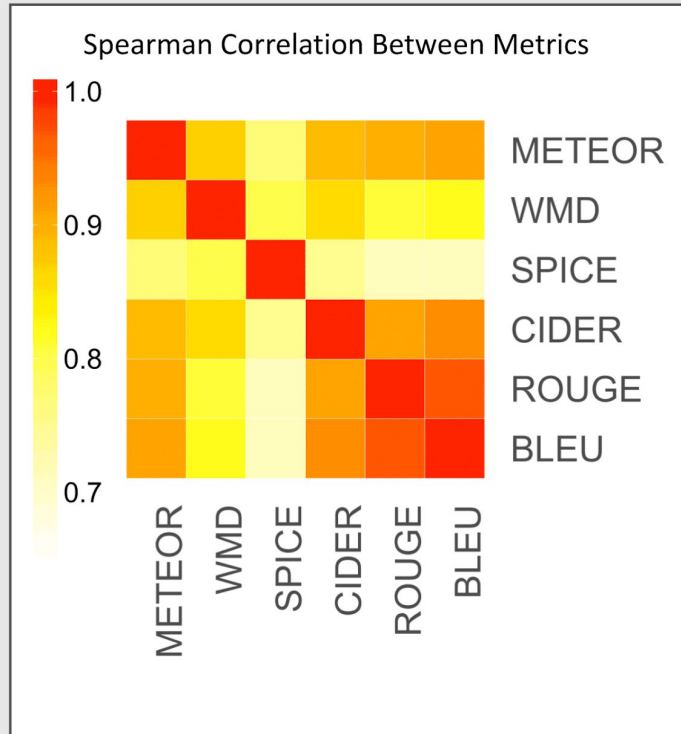
Exp 1. How correlated are the measurements provided by the metrics?

Exp 2. How significant the improvement of one metric to another?

Exp 3. Can metrics handle syntactic change?

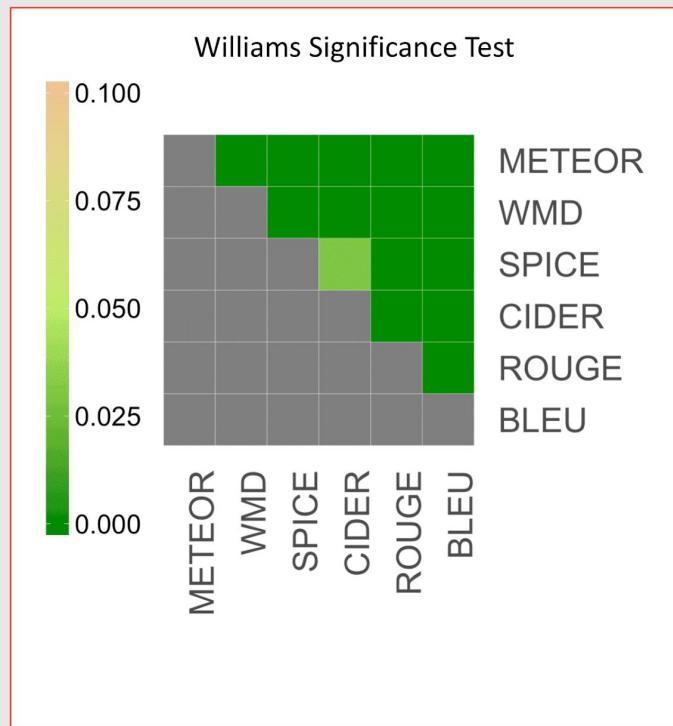
Exp 4. Can metrics handle semantic change?

## Exp 1. How correlated are the measurements provided by the metrics?



Metrics give similar responses even though they are formulated differently!

## Exp 2. How significant the improvement of one metric to another?



Improvements are significant

## Exp 3. Can metrics handle syntactic change?

Metric Values

	Description	BLEU	METEOR	ROUGE	CIDEr	SPICE	WMD
Original	A man wearing a red life jacket is sitting in a canoe on a lake	1	1	1	10	1	1
Candidate	A man wearing a life jacket is in a small boat on a lake	0.45	0.28	0.68	2.19	0.40	0.19

## Exp 3. Can metrics handle syntactic change? **Mostly no.**

Metric Values

	Description	BLEU	METEOR	ROUGE	CIDEr	SPICE	WMD
Original	A man wearing a red life jacket is sitting in a canoe on a lake	1	1	1	10	1	1
Candidate	A man wearing a life jacket is in a small boat on a lake	0.45	0.28	0.68	2.19	0.40	0.19
Synonyms	A <b>guy</b> wearing a life <b>vest</b> is in a small boat on a lake	0.20 (↓)	0.17 (↓)	0.57(↓)	0.65(↓)	0.00 (↓)	0.10(↓)
Redundancy	A man wearing a life jacket is in a small boat on a lake <b>at sunset</b>	0.45	0.28	0.66	2.01	0.36	0.18
Word order	<b>In a small boat on a lake</b> a man is wearing a life jacket	0.26 (↓)	0.26 (↓)	0.38(↓)	1.32 (↓)	0.40	0.19

## Exp 4. Can metrics handle semantic change?



Distractor Type

**Gold Caption**

A man wearing a life jacket is in a small boat on a lake with a ferry in view

**“Replace Scene”**

A man wearing a life jacket is in a small boat on **takeoff** with a ferry in view

**“Replace Person”**

**A woman in a blue shirt and headscarf** is in a small boat on a lake with a ferry in view

**“Share Person”**

**A man** is selecting a chair from a stack under a shady awning

**“Share Scene”**

**A black and brown dog is playing on the ice at the edge of a lake**

## Exp 4. Can metrics handle semantic change? **Room for improvement**

### Accuracy

Case	#Instances	BLEU	METEOR	ROUGE	CIDEr	SPICE	WMD
Replace Scene	2514	0.62	0.69	0.63	<b>0.83</b>	0.54	0.76
Replace Person	5817	0.73	0.77	0.78	0.78	0.67	<b>0.80</b>
Share Scene	2621	0.79	0.85	0.79	0.81	0.70	<b>0.87</b>
Share Person	4596	0.78	0.85	0.78	0.83	0.67	<b>0.88</b>
<b>Overall</b>	15548	0.73	0.79	0.75	0.81	0.65	<b>0.83</b>

## Conclusions

1. Metrics are lacking semantic information -> WMD

2. Metrics should be thoroughly evaluated -> Syntactically / Semantically

3. Metrics are redundant and sensitive to small changes in the sentence

4. Room for improvement for better image caption evaluation metrics