



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



Background

source

target

$$f\left(\text{source}\right) = \text{target}$$

The cat sat on the mat

?

Evaluation

Human

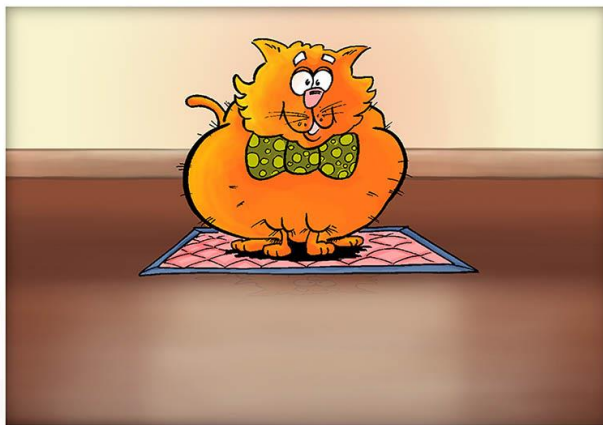


Automatic

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

Reference

{The cat sat on
the mat}



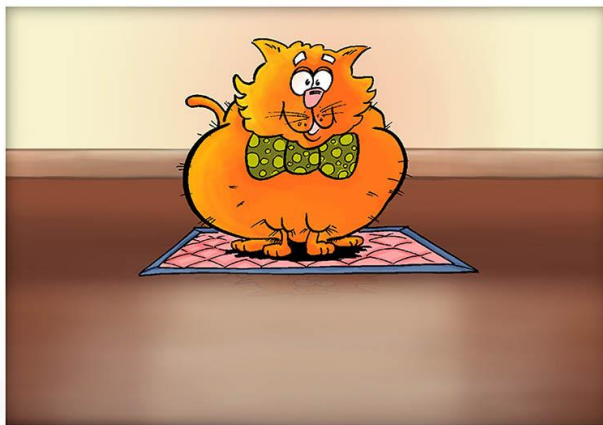
Candidate

{An orange cat
sitting on mat}

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

Reference

{The cat sat on
the mat}



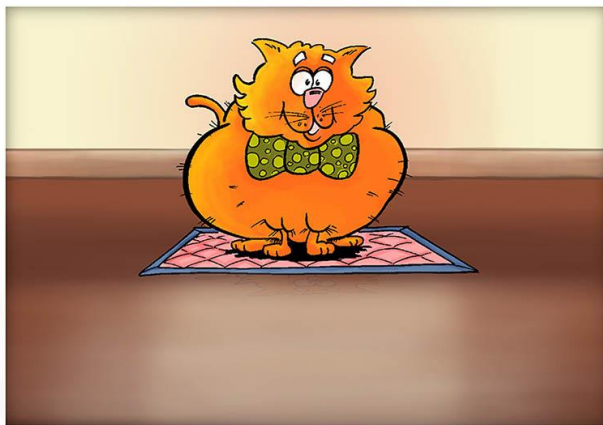
Candidate

{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

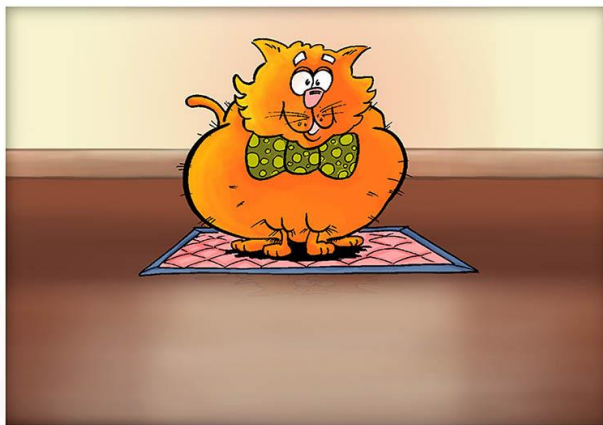
{An orange kitty
sitting on mat}

METEOR $\cong$

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

Reference

{The cat sat on
the mat}



Candidate

{An orange kitty
sitting on mat}

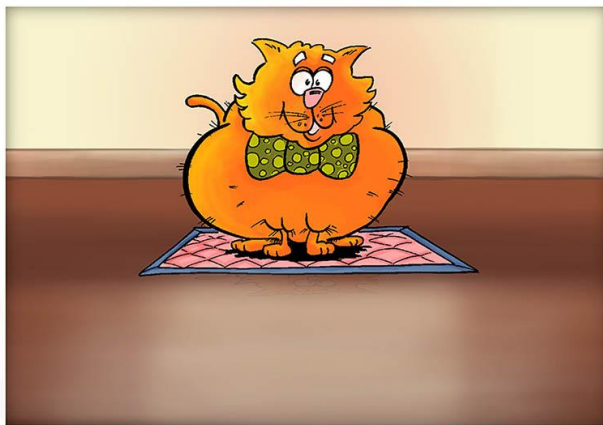
CIDEr

$f(\text{corpus}) =$

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}

CIDEr

$\approx$

$\frac{\{0.6 * \textit{kitty}, 0.2 * \textit{sit}, 0.1 * \textit{on}, 0.1 * \textit{mat}\}}$

$\{ \textit{the}, \textit{cat}, \textit{sat}, \textit{on}, \textit{the}, \textit{mat} \}$

$\times$

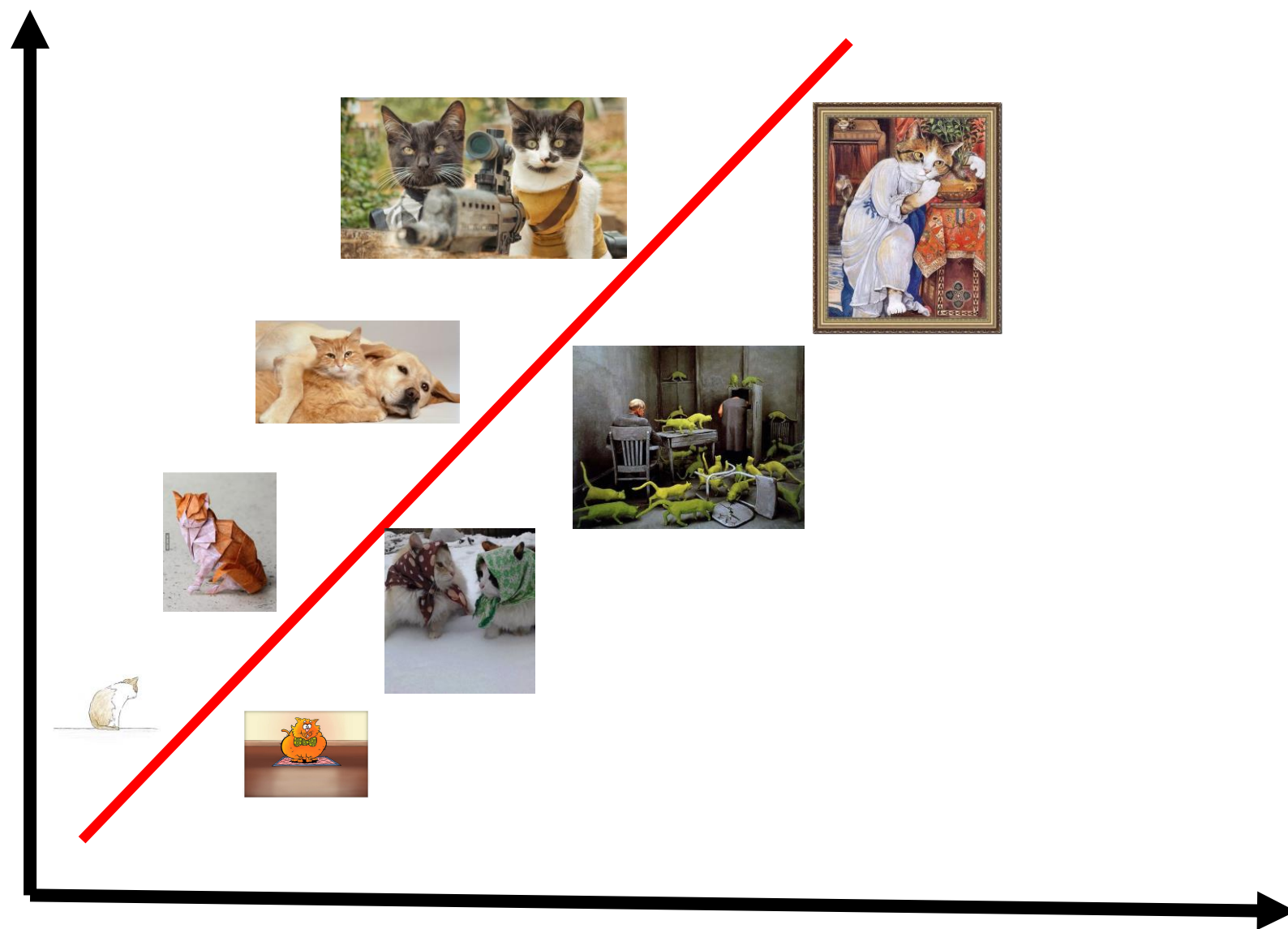
$\frac{\{0.6 * \textit{kitty}, 0.2 * \textit{sit}, 0.1 * \textit{on}, 0.1 * \textit{mat}\}}$

$\{ \textit{an}, \textit{orange}, \textit{cat}, \textit{sitting}, \textit{on}, \textit{mat} \}$

What is missing?

Metric	Proposed to Evaluate	Idea	Semantic Information
BLEU	Machine Translation	N_{gram} precision	-
ROUGE	Document Summarization	N_{gram} recall	-
METEOR	Machine Translation	N_{gram} with synonym matching	Synonym Matching
CIDEr	Image Captioning	N_{gram} with corpus re-weighting	-

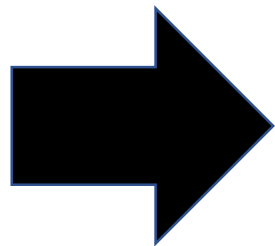
Possible Captions



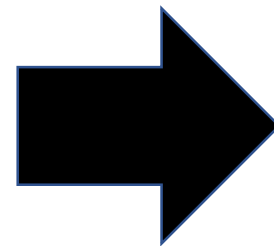
Visual Complexity

SPICE

Captions



Scene Graph



METEOR

	Reference	Candidate
<i>p</i> _{parser}	{The cat sat on the mat}	{An orange kitty sitting on mat}
<i>p</i> _{objects}	{cat, mat}	{kitty, mat}
<i>p</i> _{attributes}	{-}	{orange-kitty}
<i>p</i> _{relations}	{cat-sat, cat-on-mat}	{kitty-sit, kitty-on-mat}

Reference

{The cat sat on
the mat}



Candidate

{An orange kitty
sitting on mat}

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

What if we have a *relevant* caption, with *no* overlapping content?



Reference

A man wearing a lifevest is sitting in a canoe

Candidate 1

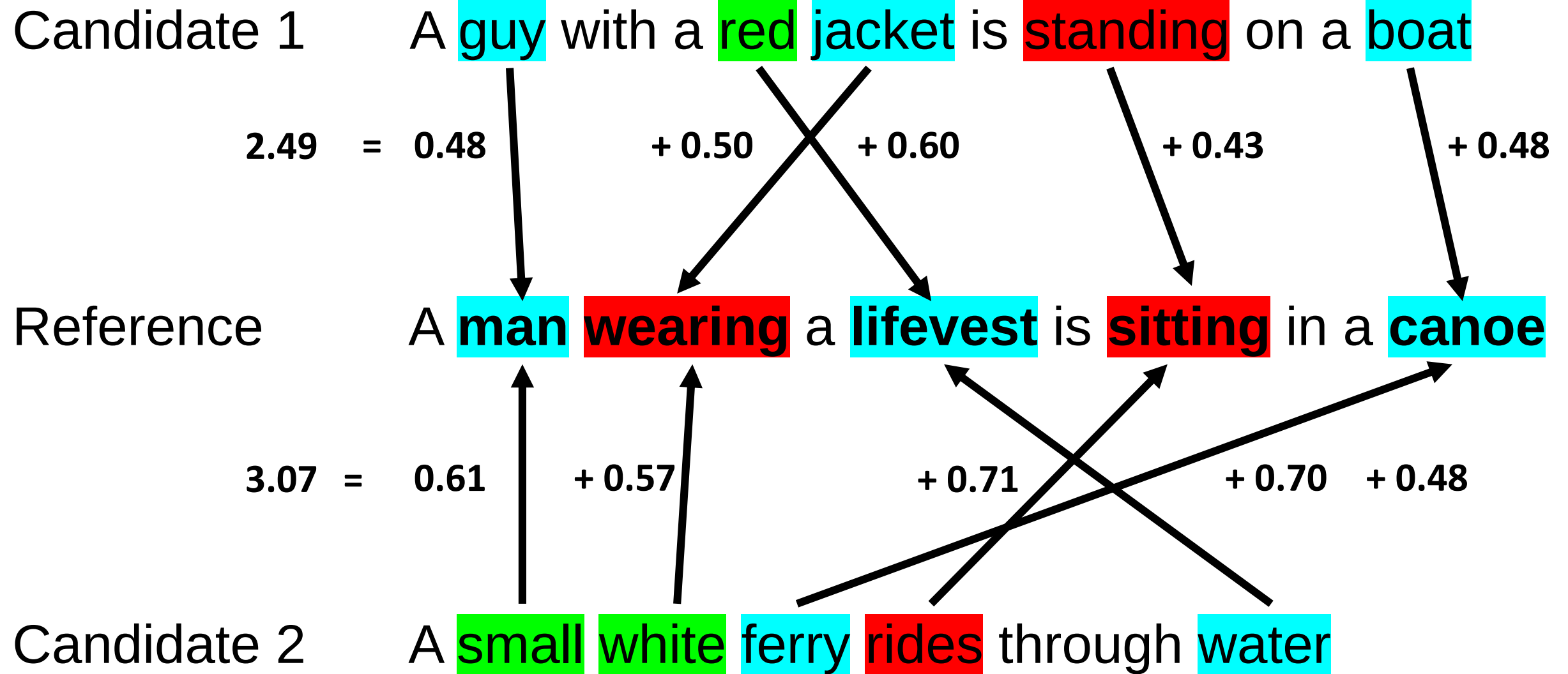
A guy with a red jacket is standing on a boat

Candidate 2

A small white ferry rides through water

Word Mover Distance (WMD)





We have *BLEU*, *METEOR*, *ROUGE*, *CIDEr*, *SPICE* and *WMD*

Which one is better?

Correlation Analysis with Williams Significance Test

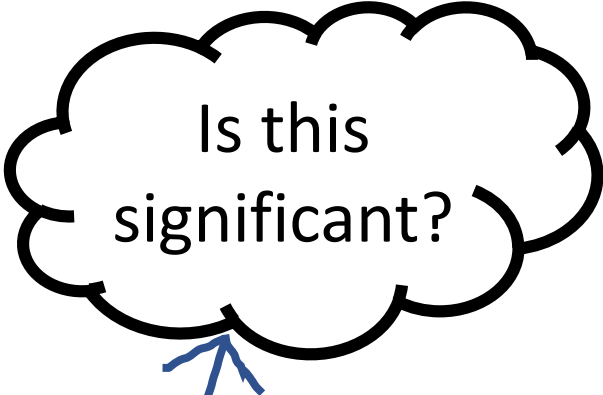
	Human Judgement	Metric 1	Metric 2
Caption 1	  	 	
Caption 2			
Caption 3		 	
Caption 4	  		 

Corr(Human Judgement, Metric 1)

Corr(Human Judgement, Metric 2)

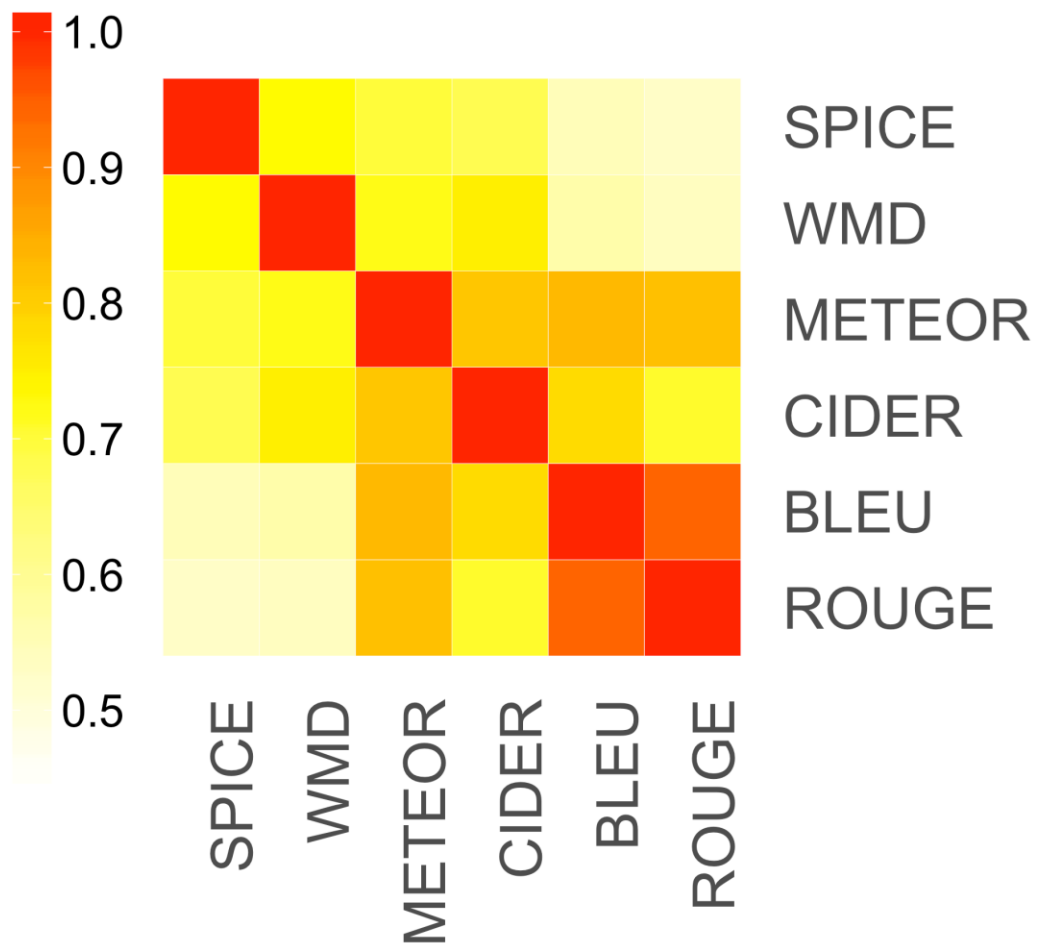
Spearman Correlation on Datasets

	Flickr-8k	Composite
WMD	0.60	0.43
SPICE	0.64	0.42
CIDEr	0.56	0.42
METEOR	0.58	0.44
BLEU	0.44	0.38
ROUGE	0.44	0.39

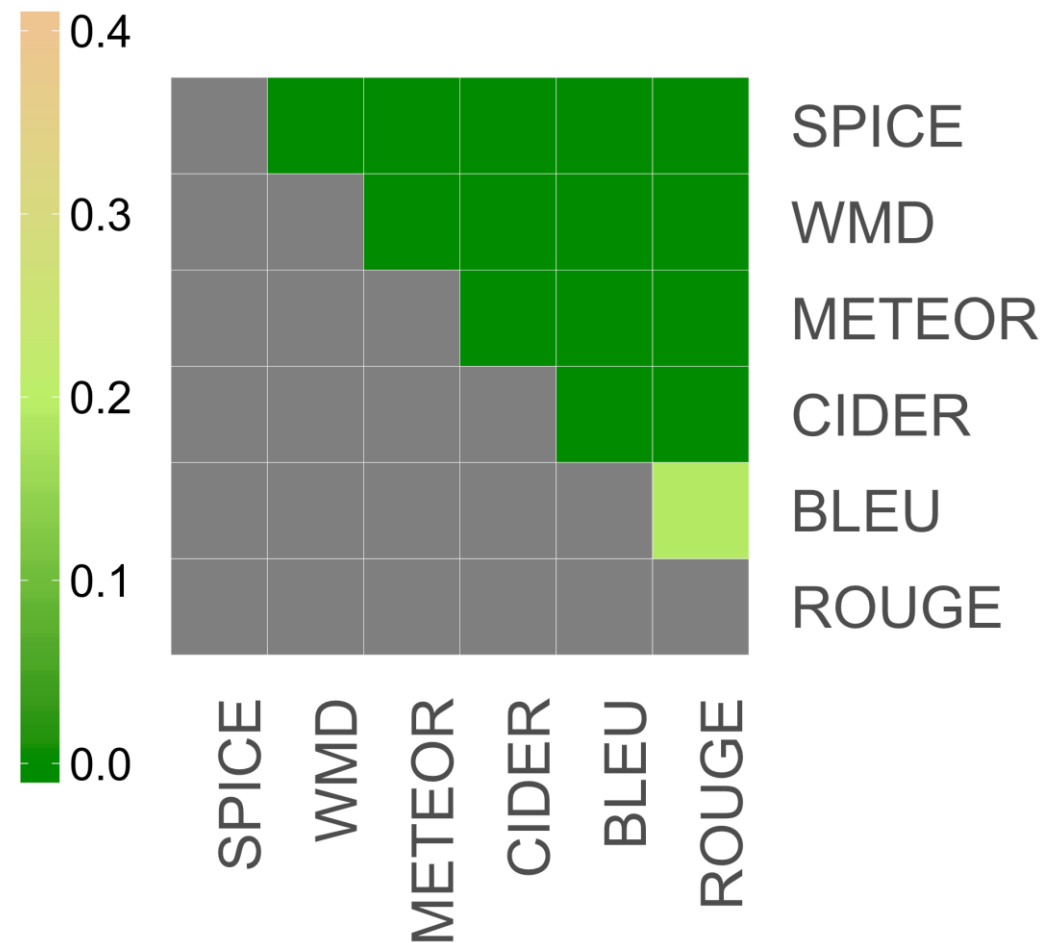


Flickr-8k	SPICE > WMD > METEOR > CIDEr > BLEU > ROUGE
Composite	METEOR > WMD > SPICE > CIDEr > ROUGE > BLEU

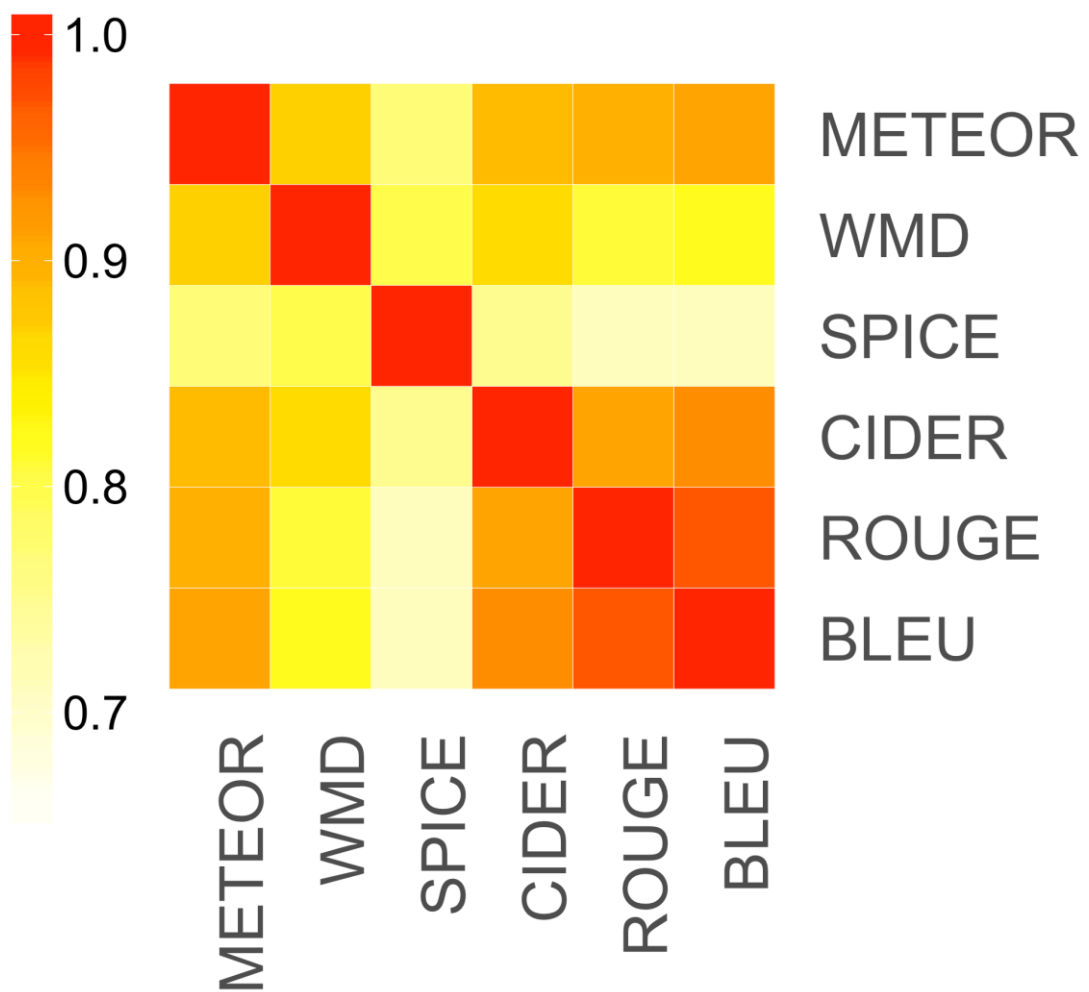
Spearman Correlation Between Metrics



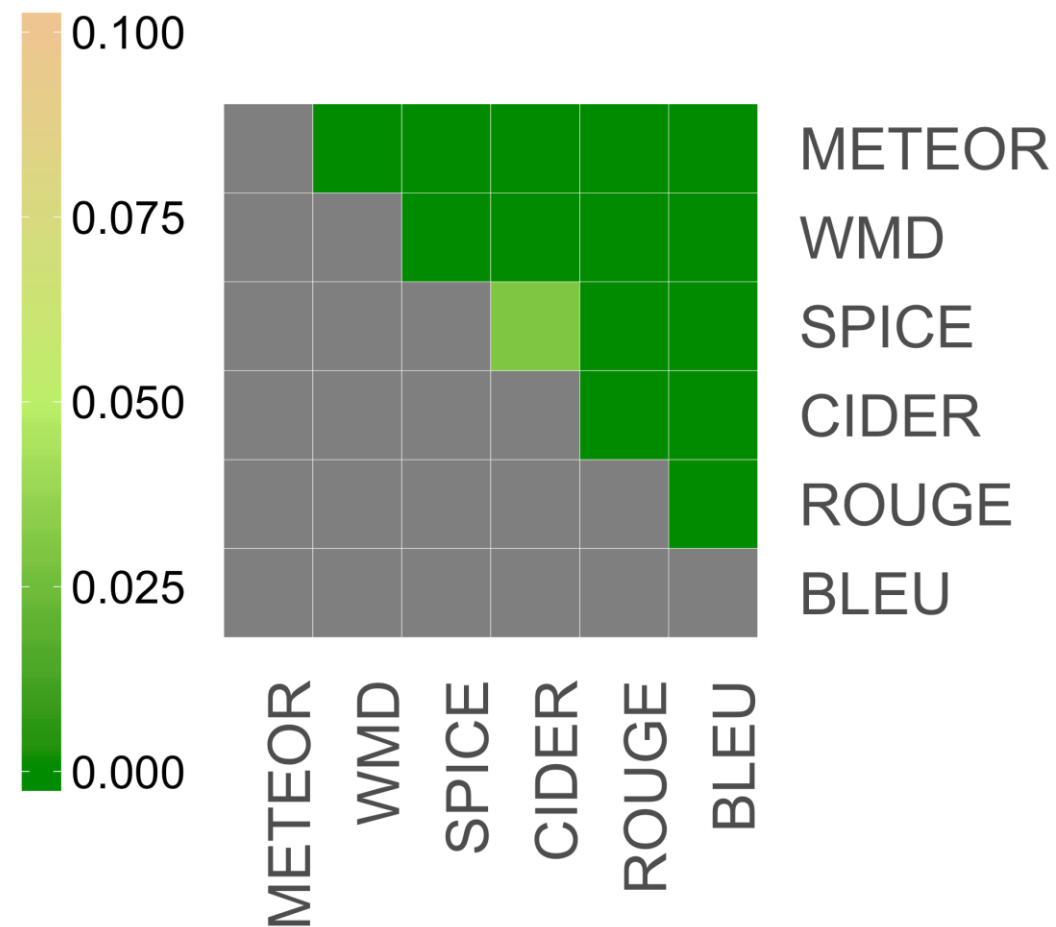
Williams Significance Test



Spearman Correlation Between Metrics



Williams Significance Test



Measuring Syntactic Properties

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

		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

		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 guy wearing a life vest is in a small boat on a lake	0.20 (↓)	0.17 (↓)	0.57(↓)	0.65(↓)	0.00 (↓)	0.10(↓)

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 guy wearing a life vest 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 at sunset	0.45	0.28	0.66	2.01	0.36	0.18

		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 guy wearing a life vest 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 at sunset	0.45	0.28	0.66	2.01	0.36	0.18
Word order	In a small boat on a lake a man is wearing a life jacket	0.26 (↓)	0.26 (↓)	0.38(↓)	1.32 (↓)	0.40	0.19

Robustness Analysis



Gold Caption

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

Distractor Type

“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 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

Accuracy

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

Conclusion

- Word Mover Distance helps as a semantic measure, and opens up opportunity to **apply semantic relevancy** literature for this task!
- Metrics tend to be effected by **small disturbances** like scene change, synonym words, or word ordering
- Is current success of image captioning metrics illusion?
- Correlation alone is not enough. It should always coupled with **Significance testing** over improvements
- There is a big room for improvement for novel metrics, and the best practices proposed in this paper can be useful to show its' contribution

References

- **[BLEU]** Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
- **[ROUGE]** Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text summarization branches out: Proceedings of the ACL-04 workshop*. Vol. 8. 2004.
- **[METEOR]** Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Vol. 29. 2005.
- **[CIDEr]** Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- **[SPICE]** Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." *European Conference on Computer Vision*. Springer International Publishing, 2016.
- **[Williams Test]** Graham, Yvette. "Improving Evaluation of Machine Translation Quality Estimation." *ACL (1)*. 2015.
- **[Robustness Corpus]** Hodosh, Micah, and Julia Hockenmaier. "Focused evaluation for image description with binary forced-choice tasks." *Workshop on Vision and Language, Annual Meeting of the Association for Computational Linguistics*. Vol. 3. 2016.
- **[Word Mover Distance]** Kusner, Matt J., et al. "From Word Embeddings To Document Distances." *ICML*. Vol. 15. 2015.