SPICE: Semantic Propositional Image Caption Evaluation

Peter Anderson[1], Basura Fernando[1], Mark Johnson[2], Stephen Gould[1]

[1] The Australian National University, Canberra, Australia | firstname.lastname@anu.edu.au
[2] Macquarie University, Sydney, Australia | mark.johnson@mq.edu.au

1. Research Questions

Can we develop an automatic image caption evalution metric that is fast and accurate? How can we understand more about the relative strengths and weaknesses of different captioning models?

2. Motivation

Existing metrics such as BLEU, METEOR, ROUGE and CIDEr are primarily sensitive to n-gram overlap. However, n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning.

'False positive'
(High n-gram similarity)

A young girl <u>standing on top of a</u> tennis court.

A giraffe <u>standing on top of a</u> green field.

'False negative'
(Low n-gram similarity)

A shiny metal pot filled with some diced veggies.

The pan on the stove has chopped vegetables in it.

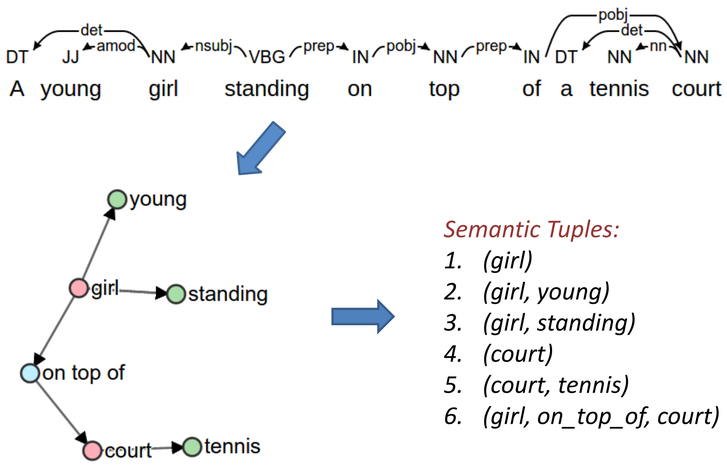
3. Approach

To overcome the limitations of existing n-gram based evaluation metrics, we focus on the semantic propositional content contained in image captions. Semantic propositions can be represented as tuples, as illustrated in the example below:

Caption: *A young girl standing on top of a tennis court.*
Semantic Propositional Content / Tuple Representation:

1. *There is a girl -> (girl)*
2. *The girl is young -> (girl, young)*
3. *The girl is standing -> (girl, standing)*
4. *There is a court -> (court)*
5. *The court is used for tennis -> (court, tennis)*
6. *The girl is on top of the court -> (girl, on_top_of, court)*

To extract these semantic propositions, reference and candidate captions are mapped through dependency parse trees [2], to semantic scene graphs [3,4] – encoding the objects (red), attributes (green) and relations (blue) present.



Semantic Tuples:

1. *(girl)*
2. *(girl, young)*
3. *(girl, standing)*
4. *(court)*
5. *(court, tennis)*
6. *(girl, on_top_of, court)*

Given a candidate caption c, a set of reference captions S, and the mapping T from captions to tuples, SPICE is calculated as an F-score over tuples in the candidate and reference scene graphs.

$$P(c,S) = \frac{|T(c) \otimes T(S)|}{|T(c)|}$$

$$R(c,S) = \frac{|T(c) \otimes T(S)|}{|T(S)|}$$

$$SPICE(c,S) = F_1(c,S) = \frac{2 \cdot P(c,S) \cdot R(c,S)}{P(c,S) + R(c,S)}$$

4. Gameability

Obvious approaches to gaming the metric have been considered:

Increasing Sentence Length:
Irrelevant content will be penalised by the F-score. Generating long, detailed and correct captions is not considered to be gaming.

Adding Synonyms:
Synonyms in the reference scene graph are collapsed, and can only be scored once.

Adding Hyponyms / Hypernyms
These are not collapsed in the reference scene graph, but they are not sufficiently common in reference captions to reward indiscriminate adding.

Note that SPICE measures recovery of objects, attributes and relations, but neglects fluency.

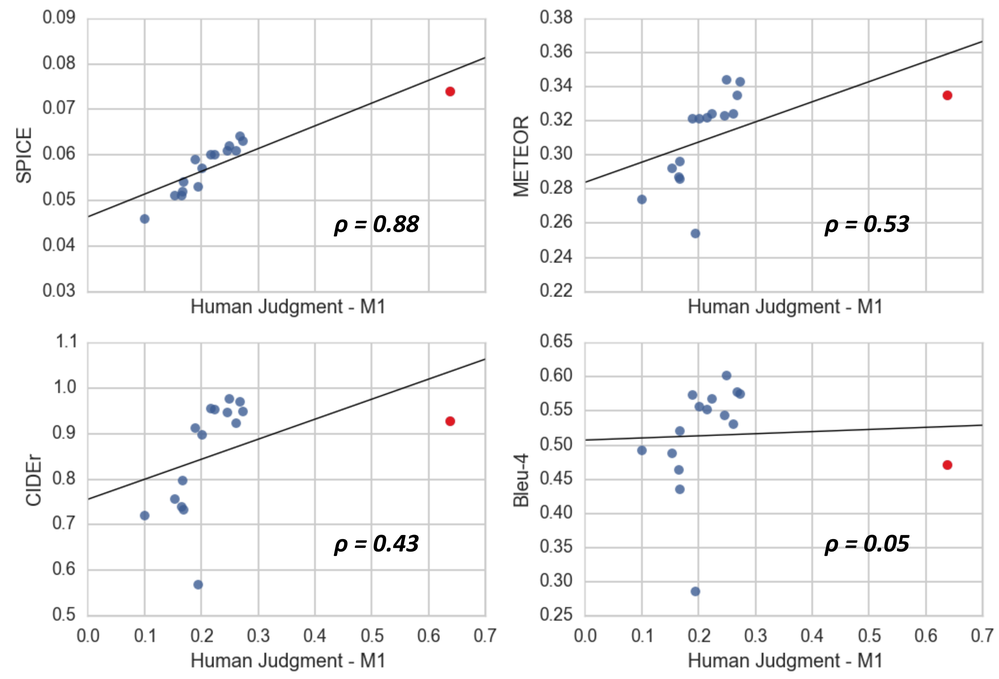
5. Re-scoring the COCO Captioning Challenge

We re-evaluated the 15 competition entries plus human captions in the 2015 COCO Captioning Challenge [1] using SPICE and other metrics. Pearson's correlation (ρ) indicates that SPICE more accurately reflects human judgment overall (M1-M2), and across each quality dimension (M3-M5).

	M1		M2		M3		M4		M5	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
Bleu-1	0.24	(0.369)	0.29	(0.271)	0.72	(0.002)	-0.54	(0.030)	0.44	(0.091)
Bleu-4	0.05	(0.862)	0.10	(0.703)	0.58	(0.018)	-0.63	(0.010)	0.30	(0.265)
ROUGE-L	0.15	(0.590)	0.20	(0.469)	0.65	(0.006)	-0.55	(0.030)	0.38	(0.142)
METEOR	0.53	(0.036)	0.57	(0.022)	0.86	(0.000)	-0.10	(0.710)	0.74	(0.001)
CIDEr	0.43	(0.097)	0.47	(0.070)	0.81	(0.000)	-0.21	(0.430)	0.65	(0.007)
SPICE-exact	0.84	(0.000)	0.86	(0.000)	0.90	(0.000)	0.39	(0.000)	0.95	(0.000)
SPICE	**0.88**	(0.000)	**0.89**	(0.000)	**0.89**	(0.000)	**0.46**	(0.070)	**0.97**	(0.000)

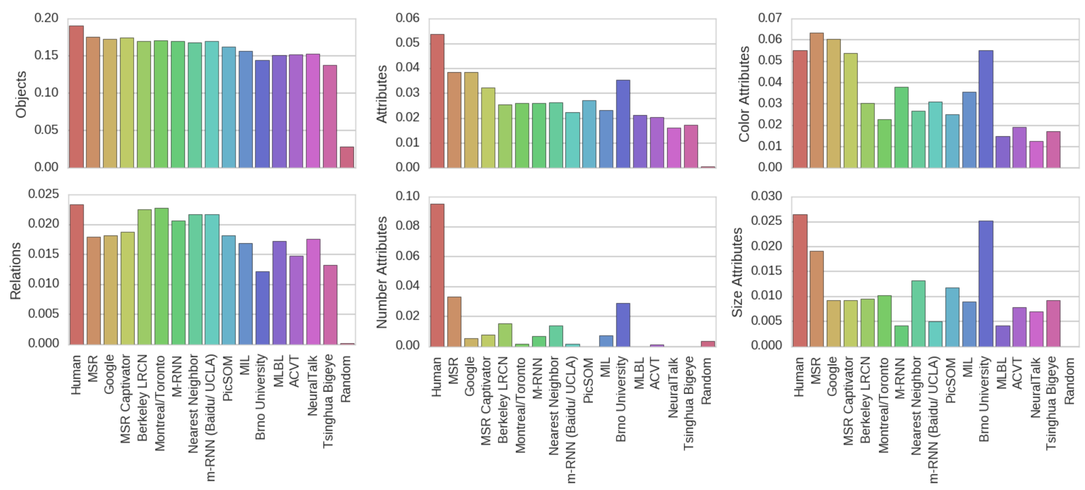
M1	Percentage of captions evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1–5 (incorrect - correct).
M4	Average detail of the captions from 1–5 (lacking details - very detailed).
M5	Percentage of captions that are similar to human description.

As illustrated in the top left plot, SPICE picks the same top 5 entries as human evaluators, and scores human-generated captions highest (shown in red).



6. Deeper Performance Evaluation

SPICE scores across various semantic proposition subcategories. These captioning models approach human performance in terms of object recovery, but appear to be less successful at recovering object attributes – particularly size and number attributes.



7. Summary

SPICE captures human judgment over image captions better than CIDEr, BLEU, METEOR and ROUGE, and enables more detailed analysis.

Acknowledgement & References

We are grateful to the COCO Consortium for agreeing to re-evaluate entries in the 2015 COCO Captioning Challenge using our SPICE code. This work was funded in part by the Australian Centre for Robotic Vision.

[1] Chen et. al. Microsoft COCO Captions: Data Collection and Evaluation Server, arXiv:1504.00325 2015
[2] Klein & Manning: Accurate Unlexicalized Parsing, ACL 2003
[3] Johnson et. al. Image Retrieval Using Scene Graphs, CVPR 2015
[4] Schuster et. al: Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval, EMNLP 2015

Code is available at the project webpage http://panderson.me/spice