MUTT: Metric Unit TesTing for Language Generation Tasks

Willie Boag, Renan Campos, Kate Saenko, Anna Rumshisky
Dept. of Computer Science, University of Massachusetts Lowell

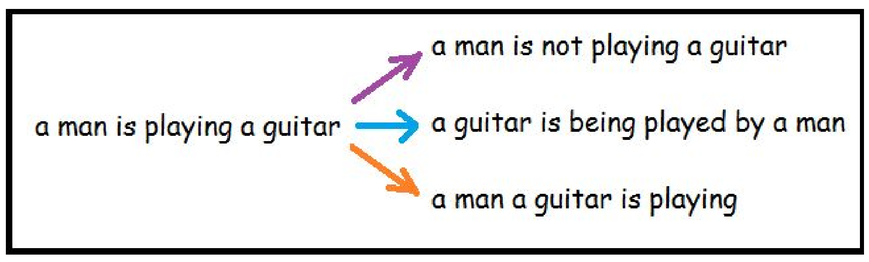
Metric Unit TesTing

Background: Language generation tasks use many evaluation metrics.

Historically: If a metric correlates well with human judgments, then we conclude that it is good.

Problem: Human judgment can be **inconsistent**. And even when it is consistent, correlation scores can still **difficult to interpret** when trying to understand specific weaknesses.

Metric Unit TesTing: A new method for metric evaluation. Measuring the effect that systematic sentence transformations have on the automatic metric scores.



Existing Metrics

BLEU - a precision-based ngram MT metric which penalizes short sentences

BADGER - a compression distance calculation after performing a series of normalization steps

CIDEr - an image captioning metric which uses a consensus-based voting of tf-idf weighted ngrams

TERp - a metric that minimizes the edit distance by stem matches, synonym matches, and phrase substitutions

METEOR - a metric that computes soft similarities between sentences by computing synonym and paraphrase scores between sentence alignments

Dataset

SICK (Sentences Involving Compositional Knowledge): a dataset for compositional distributional semantics. pairs of sentences and their human-given semantic relatedness score.

A man is holding a frog	There is no man holding a frog	2.1
The man is playing with a skull	There is no man playing with a skull	2.3
A man is driving a car	There is no man driving the car	3.6
A woman is combing her hair	There is no woman combing her hair	3.6
A man is walking outside	There is no man walking outside	4.6
A man is playing soccer	There is no man playing soccer	4.8

SICK+: Since SICK is for compositional semantics, all sentences have proper grammar. We automatically generated ungrammatical sentences (without human-estimated scores) to supplement the existing sentence pairs.

Inconsistency of Human Judgment

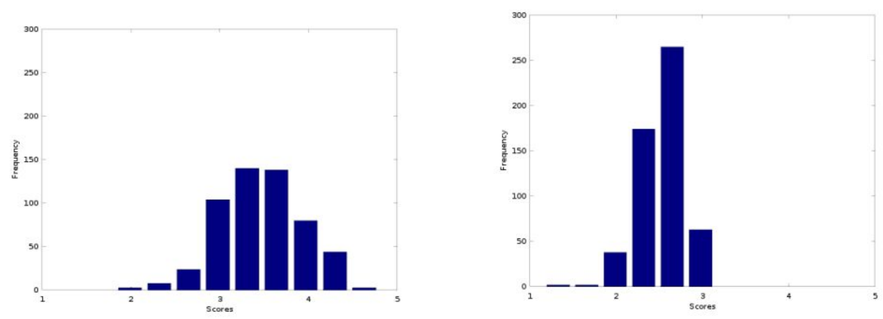


Figure 2: Human annotations for the *Negated Subject* corruption.

Figure 3: Metric predictions for the *Negated Subject* corruption.

Case Study: SICK: We examine how well human annotators are able to estimate the similarity of a sentence pair (from 1-5).

Qualitative: Figures 2 and 3 show the **wide** distribution of human-estimated labels (left) compared to the **peaky, consistent** distribution of metric-assigned labels (right).

Quantitative: Table 3 compares how strongly human-given scores ("gold") correlate with three very different automatic evaluation metric scores. **Although each metric hardly correlates with the "gold"labels, they have high correlation with one another.**

	gold	METEOR	BADGER	TERp
gold	1.00	0.09	0.03	0.07
METEOR	0.09	1.00	0.91	0.80
BADGER	0.03	0.91	1.00	0.80
TERp	0.07	0.80	0.80	1.00

Table 3: Pairwise correlation between the predictions of three evaluation metrics and the gold standard.

Method

	Meaning-Altering		
2	*negated action* (202)	"A jet is flying"	"A jet is not flying"
3	*antonym replacement* (246)	"a dog with short hair"	"a dog with long hair"
	Meaning-Preserving		
4	*active-to-passive* (238)	"A man is cutting a potato"	"A potato is being cut by a man"
6	*determiner substitution* (65)	"A cat is eating food"	"The cat is eating food"
	Fluency Disruptions		
8	*remove head from PP* (500)	"A man danced in costume"	"A man danced costume"
9	*re-order chunks* (500)	"A woman is slicing garlics"	"Is slicing garlics a woman"

Rationale: While Meaning-Altering and Fluency Disruption change a sentence's semantics, Meaning-Preserving corruptions do not.

Task: Measure the fraction of times that a given metric is able to appropriately handle a particular corruption. An accuracy of 75% would indicate that the metric is able to assign appropriate scores 3 out of 4 times.

Results and Conclusions

Unit Tests:
Corruptions that change the semantics of a sentence are "correct" if the original sentence is scored higher than the corrupted one.

Meaning-Altering
Fluency Disruption

$$s_{orig} > s_{corr}$$

Meaning-Preserving

$$\left| \frac{s_{orig} - s_{corr}}{s_{orig} + \epsilon} \right| \leq 0.15$$

Meaning-Preserving corruptions are considered "correct" if the assigned metric score of the corruption is within 15% of the original's score.

4. Active-to-Passive			
num refs	5	10	20
CIDEr	5.5	0.8	2.5
BLEU	7.6	4.6	3.8
METEOR	**23.9**	**16.0**	**13.0**
BADGER	13.4	11.3	12.2
TERp	20.6	16.4	9.7

2. Negated Action			
num refs	5	10	20
CIDEr	98.5	98.5	98.5
BLEU	97.5	97.5	98.0
METEOR	96.0	96.0	97.0
BADGER	93.6	95.5	96.5
TERp	95.5	97.0	95.0

3. Antonym Replacement			
num refs	5	10	20
CIDEr	**86.2**	**92.7**	**93.5**
BLEU	76.4	85.4	88.6
METEOR	80.9	86.6	91.5
BADGER	76.0	85.8	88.6
TERp	75.2	79.7	80.1

6. DT Substitution			
num refs	5	10	20
CIDEr	40.0	38.5	56.9
BLEU	21.5	27.7	53.8
METEOR	55.4	55.4	70.8
BADGER	**80.0**	**84.6**	**95.4**
TERp	6.2	10.8	27.7

8. Remove Head From PP			
num refs	5	10	20
CIDEr	69.5	76.8	80.8
BLEU	**63.5**	**81.3**	**87.7**
METEOR	60.6	72.9	84.2
BADGER	63.1	67.0	71.4
TERp	52.7	66.5	70.4

9. Re-order Chunks			
num refs	5	10	20
CIDEr	91.4	95.6	96.6
BLEU	83.0	91.4	94.2
METEOR	**81.2**	**89.6**	**92.4**
BADGER	95.4	96.6	97.8
TERp	91.0	93.4	93.4

Results: These tables show how well each metric is able to do (from 0% to 100%) for each corruption and a given number of reference sentences. Each entry denotes what percent of trials were "correct"

Analysis:

- **Active-to-Passive**: Accuracies decrease as metrics get mistakenly more confident in the active sentences.
- **Antonym Replacement**: CIDEr does particularly well because the misplaced antonym will "stick out like a sore thumb" with CIDEr's tf-idf weighting
- **Re-order Chunks**: METEOR gets slightly fooled because the chunks are still intact, despite being out of order