# A Study of Automatic Metrics for the Evaluation of Natural Language Explanations

Miruna Clinciu, Arash Eshghi and Helen Hastie
Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, UK

## Overview

- Explanations are a core component of human interaction, e.g. robotics, deep learning
- Strong focus on evaluation methods, common practice for NLG researchers
- Can we adopt existing NLG Metrics?

## The ExBAN Corpus

(**Ex**planations for **BA**yesian *N*etworks)
collected in a two step process:

1. **NL explanations** were produced by human subjects (84 participants)
2. In a separate study, these explanations were rated on a 7-point Likert scale, in terms of **Informativeness** and **Clarity** (97 participants, 250 explanations)

## NLG Evaluation Methods

- Human NLG Evaluation Metrics:
  - Informativeness
  - Clarity
- Automatic NLG Evaluation Metrics:
  - BLEU, ROUGE, METEOR, BERTScore & BLEURT

**ExBAN Corpus**
**Scan the QR Code**

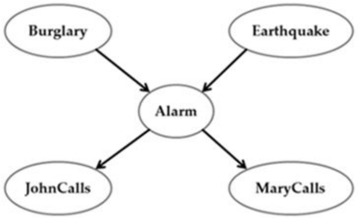


**Diagram 1**

Ref: "In the event of either burglary or earthquake the alarm will call John and Mary."

## Results: Correlation of Automatic Metrics with Human Evaluation

### Informativeness

| Metric | Diagram 1 | Diagram 2 | Diagram 3 | All Diagrams |
|---|---|---|---|---|
| BLEU-1 | 0.27 | 0.25 | 0.41* | 0.31* |
| BLEU-2 | 0.24 | 0.27 | 0.44* | 0.33* |
| BLEU-3 | 0.15 | 0.23 | 0.39 | 0.26* |
| BLEU-4 | 0.02 | 0.21 | 0.13 | 0.13 |
| SacreBleu | 0.24 | 0.30 | 0.40* | 0.30* |
| METEOR | 0.11 | -0.04 | 0.16 | 0.09 |
| Rouge-1 | 0.27 | 0.24 | 0.41* | 0.29* |
| Rouge-2 | 0.11 | 0.29 | 0.48* | 0.29* |
| Rouge-L | 0.29 | 0.28 | 0.34 | 0.29* |
| BERTScore | **0.37** | 0.21 | 0.52* | 0.37* |
| BLEURT | 0.25 | **0.38** | **0.58*** | **0.39*** |

*S*ignificance of correlation: "*" denotes p-values $< 0.05$

### Clarity

| Metric | Diagram 1 | Diagram 2 | Diagram 3 | All Diagrams |
|---|---|---|---|---|
| BLEU-1 | 0.25 | 0.09 | 0.34 | 0.24* |
| BLEU-2 | 0.24 | 0.15 | 0.41* | 0.22 |
| BLEU-3 | 0.01 | 0.10 | 0.31 | 0.14 |
| BLEU-4 | -0.01 | 0.09 | 0.18 | 0.10 |
| SacreBleu | 0.16 | 0.15 | 0.38 | 0.23 |
| METEOR | 0.17 | 0.13 | 0.30 | 0.21 |
| Rouge-1 | 0.20 | 0.11 | 0.29 | 0.20 |
| Rouge-2 | 0 | **0.24** | 0.46* | 0.22 |
| Rouge-L | 0.21 | 0.09 | 0.33 | 0.21 |
| BERTScore | **0.33** | 0.23 | 0.43* | 0.33* |
| BLEURT | 0.26 | 0.22 | **0.53*** | **0.34*** |

*S*ignificance of correlation: "*" denotes p-values $< 0.05$

- Word-overlap metrics, such as BLEU (B), METEOR (M) and ROUGE (R)
  - presented low correlation with human ratings
  - they rely on word overlap and are not invariant to paraphrases
- BERTScore (BS) and BLEURT (BRT)
  - produced higher correlation with human ratings than other metrics
  - seem to capture some relevant facts of explanations

## Good and Bad Examples of Explanations

The **alarm** is triggered by a **burglary** or an **earthquake**.

| B1 | B2 | B3 | B4 | SB | M | R1 | R2 | RL | BS | BRT | Inf. | Clar. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.19 | 0.12 | 0 | 0 | 0.05 | 0.23 | 0.25 | 0.09 | 0.12 | 0.51 | 0.52 | 7 | 7 |

Sensors = **Alarm** = prevention or ALERT.

| B1 | B2 | B3 | B4 | SB | M | R1 | R2 | RL | BS | BRT | Inf. | Clar. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.06 | 0 | 0 | 0 | 0.01 | 0.04 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

- All metrics are reasonably good at capturing and evaluating the "Bad" examples of explanations
- BLEURT (BRT) is more sensitive to Informativeness and Clarity as it captures both "Good" and "Bad" examples of explanations.
  *A larger study might be needed to show this empirically.*

## Conclusions & Future Work

- Finding accurate measures is challenging, particularly for explanations
- For future work, we plan to investigate the pragmatic and cognitive processes underlying explanations
- The ExBAN corpus and this study will inform the development of NLG algorithms for NL explanations from graphical representations.