The 16th Conference of the European Chapter
of the Association for Computational Linguistics (EACL'21)
19 to 23 of April, 2021

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

Outline

- Do NLG metrics map onto evaluation of explanations?
- Analysis of automatic metrics and whether they correlate with human judgements
- Examples of Good/Bad Explanations based on these metrics for the ExBAN corpus

Automatic Evaluation of NL Explanations

- 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? Do they capture properties of explanations?

The ExBAN Corpus

The ExBAN Corpus (Explanations for BAyesian Networks) consists of NL Explanations collected in a two step process:

1. NL explanations were produced by human subjects
   - Total number of participants: 84
2. In a separate study, these explanations were rated on a 7-point Likert scale, in terms of Informativeness and Clarity
   - Total number of explanations: 250
   - Total number of participants: 97
   - Total number of ratings: 2910

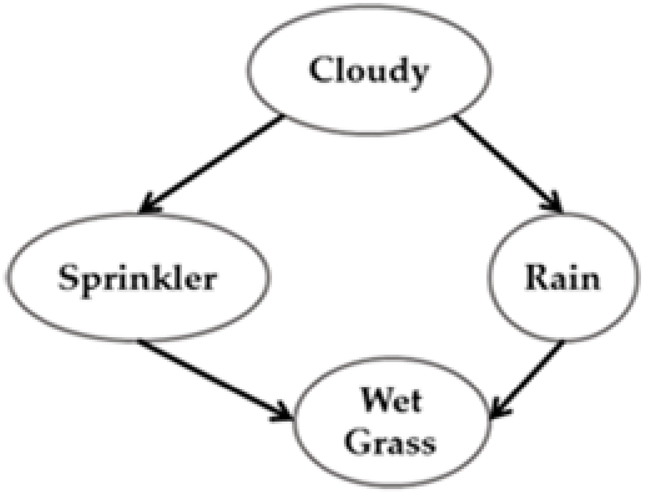

Diagram 2

Ref: "If it gets cloudy, it can rain or the sprinkler may get activated. Whenever it rains or the sprinkler gets activated, the grass gets wet."

NLG Evaluation Methods

- Human NLG Evaluation Metrics:
  - Informativeness (Novikova et al., 2018)
  - Clarity (Belz and Kow, 2009; van der Lee et al., 2017)
- Automatic NLG Evaluation Metrics: BLEU, ROUGE, METEOR, BERTScore and BLEURT

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***

Significance 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***

Significance of correlation: "*" denotes p-values < 0.05

Results: Correlation of Automatic Metrics with Human Evaluation

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)

- outperformed other metrics
- produced higher correlation with human ratings than other metrics
- seem to capture some relevant facts of explanations

Good and Bad Examples of Explanations

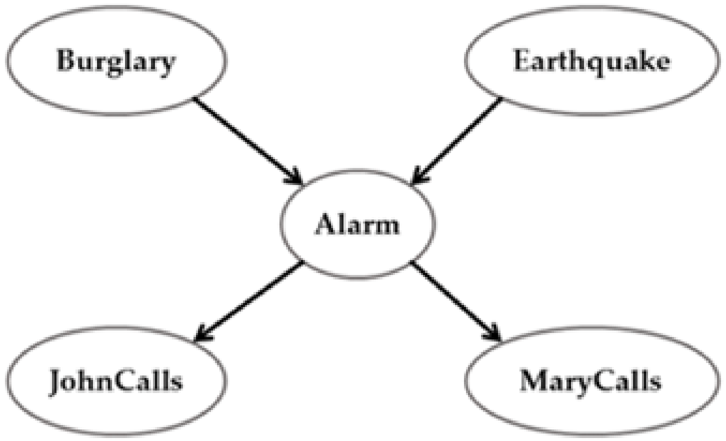

Diagram 1

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

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

The words that represents the nodes of a BN graphical model representation, are **bolded.**

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Conclusions and Future Work

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

Thank you for your attention!

ExBAN Corpus

Scan the QR Code

EDINBURGH CENTRE FOR ROBOTICS

Innovation Ready

Bibliography

Clinciu, M. A., & Hastie, H. F. (2019). A survey of explainable AI terminology. In *NL4XAI 2019 - 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, Proceedings of the Workshop* (pp. 8–13). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/w19-8403

Hastie, H., & Belz, A. (2014). A comparative evaluation methodology for NLG in interactive systems. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (pp. 4004–4011). European Language Resources Association (ELRA).

Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *ArXiv, abs/1703.09902*.

Novikova, J., Dušek, O., & Rieser, V. (2018). RankME: Reliable human ratings for natural language generation. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 2, pp. 72–78). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/n18-2012

Belz, A., & Kow, E. (2009). System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG 2009* (pp. 16–24). Association for Computational Linguistics (ACL). https://doi.org/10.3115/1610195.1610198

Van Der Lee, C., Krahmer, E., & Wubben, S. (2017). PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *INLG 2017 - 10th International Natural Language Generation Conference, Proceedings of the Conference* (pp. 95–104). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/w17-3513