I don't understand! Evaluation Methods for Natural Language Explanations

Miruna Clinciu[1,2], Arash Eshghi[1], and Helen Hastie[1]

[1] Heriot−Watt University, Edinburgh, UK
[2] University of Edinburgh, Edinburgh, UK
{mc191,a.eshghi,h.hastie}@hw.ac.uk

Abstract. Explainability of intelligent systems is key for future adoption. While much work is ongoing with regards to developing methods of explaining complex opaque systems, there is little current work on evaluating how effective these explanations are, in particular with respect to the user's understanding. Natural language (NL) explanations can be seen as an intuitive channel between humans and artificial intelligence systems, in particular for enhancing transparency. This paper presents existing work on how evaluation methods from the field of Natural Language Generation (NLG) can be mapped onto NL explanations. Also, we present a preliminary investigation into the relationship between linguistic features and human evaluation, using a dataset of NL explanations derived from Bayesian Networks.

Keywords: Explanations · Evaluation · Natural Language

[3]

1 Introduction

The rapid advance of Artificial Intelligence poses some fundamental ethical and social concerns, where providing the right explanations for instilling transparency in AI systems represents a main topic of discussion. An intuitive medium to provide explanations is through natural language, and with recent regulations comes an increasing need to provide evaluation methods for natural language explanations that will help us to assess the quality of those explanations in relation to the system they explain. This need for evaluating explanations has been further validated by studies from social sciences and psychology [2, 8, 11, 17].

Particular questions arise around the amount of information needed to explain but not overload the user, the lexical choice for matching the user's understanding and expertise level, as well as the linguistic style adopted. Attributes such as informativeness, clarity, coherence, readability and effectiveness have

been linked to human evaluation dimensions frequently used in the field of Natural Language Generation (NLG) [2, 17]. Considering the strong focus of the NLG researchers on evaluating natural language, we propose that mapping existing NLG methods onto NL explanations can provide insights into the definition of a good explanation [3].

To get a better understanding of how we can define what makes an effective NL explanation, we designed and gathered the ExBAN Corpus (Explanations from Bayesian Networks). This corpus provides NL explanations for a set of Bayesian Networks, mainly motivated by the fact that Bayesian Networks are frequently used for the detection of anomalies in data and can approximate deep learning models. They also allow us to sense-check our explanation evaluation techniques as they are reasonably easy to understand for the non-expert user.

This paper presents current work into the evaluation of NL explanations , but also includes preliminary new linguistic analysis. The paper is thus structured into four parts: (1) we introduce the ExBAN corpus; (2) we present how automatic and human evaluation metrics from the field of NLG can be mapped onto NL explanations; (3) we present an analysis on how linguistic features correlate with human evaluation metrics; and (4) finally, we discuss how evaluation methods can capture the quality of NL explanations. Further details of this work can be found in [3].

2 ExBAN Corpus

Existing datasets of explanations have enabled significant progress in the way that explanations provide transparency of machine learning algorithms. However, less attention has been paid to methods to explain structured data, such as Bayesian Networks. Bayesian Networks have "the ability to cover any model with a probabilistic interpretation including supervised, unsupervised, and reinforcement learning (including deep learning)" [16]. Also, their graphical representation can be used for extracting information [7]. The ExBAN corpus is used here for evaluation, but it could also be used for training models for generating natural language explanations from graphical models such as Bayes Nets and other structured data more broadly.

2.1 ExBAN Corpus Desciption

Definition: ExBN: A Corpus of Natural Language Explanations for three Bayesian Networks's graphical representations (see Figure 1).
Purpose: Possible application areas for the corpus: explainable AI, general artificial intelligence, academic linguistic research, natural language processing.

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 (a total number of 84 participants)

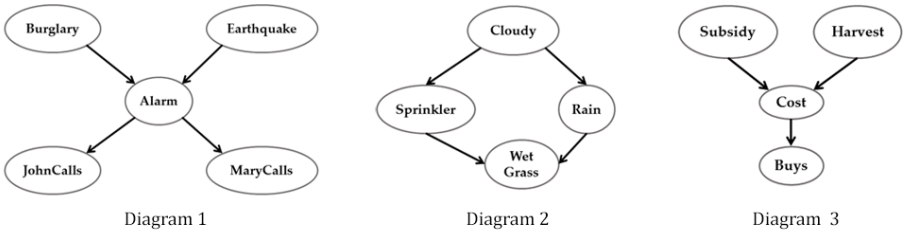

Fig. 1. Annotated Diagrams: where **Diagram 1** represents a typical Bayesian Network, **Diagram 2** represents a multiply-connected network and **Diagram 3** represents a simple network with discrete variables (Subsidy and Buys) and continuous variables (Harvest and Cost) [3].

2. In a separate study, these explanations were rated on a 7-point Likert scale, in terms of informativeness and clarity (a total number of 250 explanations, rated by 97 participants; each explanation was rated by minimum of 3 participants)

3 NLG Evaluation Methods

Models trained iteratively with large amounts of data are particularly hard to evaluate in a cost-effective and timely manner. Therefore, creating automatic methods for evaluating NLG systems that can capture the human-likeness of the generated output is essential.

Human Evaluation. Explanations should be clear and easily understood by users, providing the right information in order to create better communication [6, 9]. We focus on two dimensions: informativeness and clarity.

Automatic Evaluation. Here we describe automatic metrics used in the field of NLG evaluation and selected for this study, specifically: 1) word-based (untrained) metrics such as BLEU, METEOR and ROUGE, and 2) pre-trained metrics, such as BERTScore and BLEURT. Here, we briefly describe each in turn:

- **BLEU** [12] is a widely used metric in the field of NLG (borrowed from Machine Translation (MT)) that compares n-grams of a candidate text (e.g. that generated by algorithms) with the n-grams of a reference text. The number of matches defines the goodness of the candidate text.
- **SacreBLEU** was proposed by [13] as a new version of BLEU that calculates scores on the detokenized text by applying its own metric-internal preprocessing.
- **METEOR** was created in order to address the weaknesses of BLEU; METEOR evaluates generated text by computing a score based on explicit word-to-word matches between a candidate and a reference. When using multiple

references, the candidate text is scored against each reference, and the best score is reported.
- **ROUGE** [10] evaluates n-gram overlap of the generated text (candidate) with a reference.
- **ROUGE-L** computes the longest common subsequence (LCS) between a pair of sentences [14].
- **BERTScore** [18] is a token-level matching metric with pre-trained contextual embeddings using BERT [5] that matches words in candidate and reference sentences using cosine similarity.
- **BLEURT** [15] is a text generation metric based on BERT, pre-trained on synthetic data; it uses "random perturbations of Wikipedia sentences augmented with a diverse set of lexical and semantic-level supervision signals". BLEURT uses a collection of metrics and models from prior work, including BLEU and ROUGE.

3.1 Correlation of Automatic Metrics with Human Evaluation

In order to investigate the degree to which automatic metrics for NLG can capture the quality of NL explanations [3], we ran a correlation analysis with automatic metrics with human judgements. As shown in Figure 2, we can draw the following conclusions:
- Word-overlap metrics such as BLEU (n = 1,2,3,4), METEOR and ROUGE (n = 1,2) presented low correlation with human ratings. This might be due to certain limitations, such as the fact that they rely on word overlap and are not invariant to paraphrases.
- BERTScore and BLEURT outperformed other metrics and produced higher correlation with human ratings than other metrics on all diagrams. These metrics might capture some relevant facts of explanations, as word representations are dynamically informed by the words around them.

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

Fig. 2. Highest absolute Spearman correlation between automatic evaluation metrics and human ratings for informativeness and clarity, where the bold font represents the highest correlation coefficient obtained by an automatic evaluation metric

According to human evaluation scores for informativeness and clarity, in Figure 3, we present examples of explanations with high scores for informativeness and clarity ("Good" Examples of Explanation) and with low scores for informativeness and clarity ("Bad" Examples of Explanation). As observed, all automatic metrics are reasonably good at capturing and evaluating the "Bad" examples of explanations. Also, we can see that only BLEURT (BRT) is more sensitive to capturing informativeness and clarity, for both examples.

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

Fig. 3. Examples of Good and Bad Explanations

4 Linguistic Features

We extracted a number of linguistic features presented in Table 1 to explore if there were any linguistic constructs that were found in our dataset and that mapped to good or bad explanations. For example, complex syntactic construction might lead to difficulty in understanding for the user, as reflected in the Height_tree and Length_tree features. Other features given in table 1 were motivated by similar studies correlating features with user ratings for surface realisation in NLG [4], and for social psychology [1].

Table 1. Linguistic features

Total_words	total number of words
Sentence_Length	average sentence length
Nr_Nouns	number of nouns per explanation (NN - singular common nouns, NNS - plural common nouns, NNP - proper noun)
WDT	number of wh-determiners which
CC	number of coordinating conjunctions
Avg_tfidf	average tf-idf score of content words
Height_tree	depth of syntactic embedding
Length_tree	the number of children it has

We mapped the linguistic features in Table 1 to human evaluation metrics (informativeness and clarity) to see if there was any correlation between these features and the quality of the explanation, as rated by humans.

Our preliminary analysis shows some trends, but more investigation is needed to confirm these. We calculated the Spearman's correlation coefficient between the linquistic features and human evaluation ratings score, for both informativeness and clarity on a sample of 166 datapoints. With regards informativeness, the sentence length ($r = 0.29$) and the number of nouns ($r = 0.36$) presents weak correlatation with informativeness, as well as the number of coordinating conjunctions ($r = 0.23$).

Linguistic features do not seem to capture well the level of clarity of a sentence as no correlation was found in this regard. This is perhaps because clarity is multi-dimensional and implies more than lexical-syntactic relationships, including other factors such as causality, common sense and general knowledge.

5 Conclusions

Finding accurate automatic measures is challenging, particularly for explanations, as the pragmatic and cognitive processes underlying explanations, such as reasoning, causality, and common sense, might not be captured. In our study, the embedding-based metrics perform better than the word-overlap based ones, but we would recommend a larger study to show this empirically. Future work would involve examining the effectiveness of automatic metrics across a wider variety of explanation tasks and datasets. Finally, the next step is to use this work to automatically generate natural language explanations from structured data such as Bayes Nets, and this work contributes towards ensuring the quality of such explanations.

Acknowledgments

This work was supported by the EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems at Heriot-Watt University and the University of Edinburgh. Clinciu's PhD is funded by Schlumberger Cambridge Research Limited (EP/L016834/1, 2018-2021). This work was also supported by the EPSRC ORCA Hub (EP/R026173/1, 2017-2021) and UKRI Trustworthy Autonomous Systems Node on Trust (EP/V026682/1, 2020-2024).

Bibliography

[1] Abelson, R.P., Leddo, J., Gross, P.H.: The Strength of Conjunctive Explanations. Personality and Social Psychology Bulletin **13**(2) (1987). https://doi.org/10.1177/0146167287132001

[2] Clinciu, M., Hastie, H.: Let's Evaluate Explanations! HRI 2020 Workshop on Test Methods and Metrics for Effective HRI in Real World Human-Robot Teams (2020)

[3] Clinciu, M.A., Eshghi, A., Hastie, H.: A study of automatic metrics for the evaluation of natural language explanations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2376–2387. Association for Computational Linguistics, Online (Apr 2021), https://www.aclweb.org/anthology/2021.eacl-main.202

[4] Dethlefs, N., Cuayáhuitl, H., Hastie, H., Rieser, V., Lemon, O.: Cluster-based prediction of user ratings for stylistic surface realisation. In: 14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014. pp. 702–711. Association for Computational Linguistics (ACL) (2014)

[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. vol. 1, pp. 4171–4186. Association for Computational Linguistics (ACL) (2019)

[6] Hastie, H.: Metrics and Evaluation of Spoken Dialogue Systems. In: Data-Driven Methods for Adaptive Spoken Dialogue Systems. Springer Publishing Company, Incorporated (2012). https://doi.org/10.1007/978-1-4614-4803-7_7

[7] Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., Hajishirzi, H.: Text generation from knowledge graphs with graph transformers. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (2019)

[8] Leake, D.B.: Evaluating Explanations. Psychology Press (feb 2014). https://doi.org/10.4324/9781315807072

[9] Lemon, O., Pietquin, O.: Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces. Springer Publishing Company, Incorporated (2012)

[10] Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew. Information Sciences Institute **34**(12) (1971). https://doi.org/10.1253/jcj.34.1213

[11] Mohseni, S., Zarei, N., Ragan, E.D.: A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. ACM Transactions on Interactive Intelligent Systems (2018)
[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.j., Heights, Y.: IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation. Science **22176**, 1–10 (2001). https://doi.org/10.3115/1073083.1073135, http://dl.acm.org/citation.cfm?id=1073135
[13] Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6319, https://www.aclweb.org/anthology/W18-6319
[14] Schluter, N.: The limits of automatic summarisation according to ROUGE. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. vol. 2 (2017). https://doi.org/10.18653/v1/e17-2007
[15] Sellam, T., Das, D., Parikh, A.: BLEURT: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.704, https://www.aclweb.org/anthology/2020.acl-main.704
[16] Yang, S.C.H., Shafto, P.: Explainable Artificial Intelligence via Bayesian Teaching. In: Neural Information Processing Systems Workshop: Teaching Machines, Robots, and Humans (2017)
[17] Zemla, J.C., Sloman, S., Bechlivanidis, C., Lagnado, D.A.: Evaluating everyday explanations. Psychonomic Bulletin & Review **24**(5), 1488–1500 (Oct 2017). https://doi.org/10.3758/s13423-017-1258-z, https://doi.org/10.3758/s13423-017-1258-z
[18] Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkeHuCVFDr