A study of automatic metrics for the evaluation of natural language explanations

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

Miruna-Adriana Clinciu
Edinburgh Centre for Robotics
Heriot-Watt University
University of Edinburgh
mc191@hw.ac.uk

Arash Eshghi
Heriot-Watt University
Edinburgh, United Kingdom
a.eshghi@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
h.hastie@hw.ac.uk

Abstract

As transparency becomes key for robotics and AI, it will be necessary to evaluate the methods through which transparency is provided, including automatically generated natural language (NL) explanations. Here, we explore parallels between the generation of such explanations and the much-studied field of evaluation of Natural Language Generation (NLG). Specifically, we investigate which of the NLG evaluation measures map well to explanations. We present the ExBAN corpus: a crowd-sourced corpus of NL explanations for Bayesian Networks. We run correlations comparing human subjective ratings with NLG automatic measures. We find that embedding-based automatic NLG evaluation methods, such as BERTScore and BLEURT, have a higher correlation with human ratings, compared to word-overlap metrics, such as BLEU and ROUGE. This work has implications for Explainable AI and transparent robotic and autonomous systems.

1 Introduction

The machine learning models and algorithms underlying today's AI and robotic systems are increasingly complex with their internal operations and decision-making processes ever more opaque. This opacity is not just an issue for the end-user, but also the creators and analysts of these systems. As we move towards building safer and more ethical systems, this lack of transparency needs to be addressed. One key trait of a transparent system is its ability to be able to *explain* its deductions and articulate the reasons for its actions in Natural Language (NL). As the area of Explainable AI (XAI) grows and is mandated (cf. the EU General Data Protection Regulation's "right to explanation" (Commission, 2018) and standardisation (cf. IEEE forthcoming standard on Transparency (P7001)), it has become ever more important to be able to evaluate the quality of the NL explanations themselves, as well as the AI algorithms they explain. Furthermore, the importance of evaluating explanations has been emphasised by researchers within the social cognitive sciences (Leake, 2014; Zemla et al., 2017; Doshi-Velez and Kim, 2017). To date, explanations have mostly been evaluated by collecting human judgements, which is both time-consuming and costly. Here, we view generating explanations as a special case of Natural Language Generation (NLG), and so we explore mapping existing automatic evaluation methods for NLG onto explanations. We study whether general, domain-independent evaluation metrics within NLG are sensitive enough to capture the peculiarities inherent in NL explanations (Kumar and Talukdar, 2020), such as causality; or whether NL explanations constitute a sui-generis category, thus requiring their own automatic evaluation methods and criteria.

In this paper, we firstly present the ExBAN dataset: a corpus of NL explanations generated by crowd-sourced participants presented with the task of explaining simple Bayesian Network (BN) graphical representations. These explanations were subsequently rated for *Clarity* and *Informativeness*, two subjective ratings previously used for NLG evaluations (Gatt and Krahmer, 2018; Howcroft et al., 2020). The motivation behind using BN is that they are reasonably easy to interpret, are frequently used for the detection of anomalies in the data (Tashman et al., 2020; Saqaeeyan et al., 2020; Metelli and Heard, 2019; Mascaro et al., 2014), and have been used to approximate deep learning methods (Riquelme et al., 2018; Gal and Ghahramani, 2016), which we could, in turn, explain in Natural Language.

Secondly, we explore a wide range of automatic measures commonly used for evaluating NLG to

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2376–2387
April 19 - 23, 2021.

understand if they capture the human-assessed quality of the corpus explanations. We then go on to discuss their strengths and weaknesses through quantitative and qualitative analysis.

Our contributions are thus as follows: (1) a new corpus of natural language explanations generated by humans, who are asked to interpret Bayesian Network graphical representations, accompanied by subjective quality ratings of these explanations. This corpus can be used in various application areas including Explainable AI, general Artificial Intelligence, linguistics and NLP; (2) a study of methods for evaluating explanations through automatic measures that reflect human judgements; and (3) qualitative discussion into these metrics' sensitivity by examining specific explanations varying on the Informativeness/Clarity scales.

2 Related Work

Explanations are a core component of human interaction (Scalise et al., 2017; Krening et al., 2017; Madumal et al., 2019). In the context of Machine Learning (ML), explanations should articulate the decision-making process of an ML model explicitly, in a language familiar to people as communicators (De Graaf and Malle, 2017; Miller, 2018). According to Plumb et al. (2018), three of the most common types of explanation are: (1) *global explanations*, which describe the overall behaviour of the entire model (Arya et al., 2019); (2) *local explanations*, commonly taking the form of counterfactuals (Sokol and Flach, 2019) that describe why particular events happened (known also as "everyday explanations"); and (3) *example-based explanations* that present examples from the training set to explain algorithmic behaviour (Cai et al., 2019).

Recently, various explanation systems provide different types of explanations for AI systems: the LIME method visually explains how sampling and local model training works by using local interpretable model-agnostic explanations (Ribeiro et al., 2016); MAPLE can provide feedback for all three types of explanations: example-based, local and global explanations (Plumb et al., 2018); CLEAR explains a single prediction by using local explanations that include statements of key counterfactual cases (White and d'Avila Garcez, 2019). Whilst these techniques and tools gain some ground in explaining deep machine learning, the explanations they provide are not necessarily aimed at the (non-expert) end-user and so are not always intuitive.

NLG has traditionally been broken down into "what" to say (content selection) and "how" to say it (surface realisation) and can draw parallels with Natural Language explanations. In particular, it is important to gauge how much content or how many reasons to present to the user, to inform them fully without overloading them. For example, prior work has shown that people prefer shorter explanations that offer only sufficient detail to be considered useful (Harbers et al., 2009; Yuan et al., 2011).

According to Miller et al. (2017), how explainers generate and select explanations depends on so-called pragmatic influences of causes, and they found that people seem to prefer simpler and more general explanations. Similarly, Lombrozo (2007) notes that simplicity and generality might be the key to evaluating explanations. This was partly the case described in (Chiyah Garcia et al., 2018), but here the users were experts and preferred to be given all possible reasons but as precise and brief as possible. It is clear from these prior works that explanations have to be evaluated in the context of the scenario, prior knowledge and preferences of the explainee, and the intent and goals of the explainer. These could be, for example, establishing trust (Miller et al., 2017), agreement, satisfaction, or acceptance of the explanation and the system (Gregor and Benbasat, 1999).

Somewhat analogous to auto-generated explanations are the fields of summarisation of text (Tourigny and Capus, 1998; Deutch et al., 2016) and Question-Answering (Dali et al., 2009; Xu et al., 2017; Lamm et al., 2020). This is because they provide users (expert and lay users) with various forms of summaries (visual or textual) and answers containing explanations to enable them to have a better understanding of content.

Summarisation methods and sentence compression techniques can help to build comprehensive explanations (Winatmoko and Khodra, 2013). With regards to evaluating these summarisation methods, Xu et al. (2020) proposed an evaluation metric that weighted the facts present in the source document according to the facts selected by a human-written (natural language) summary, by using contextual embeddings. This evaluation of text accuracy

is indeed related to explanations because any explanation must contain enough statements to support decision-making and understanding. These statements should be accurate and true.

The growing interest in the AI community to investigate the potential of NL explanations for bridging the gap between AI and HCI has resulted in an increasing number of NL explanations datasets. The ELI5 dataset[1] (Fan et al., 2019) is composed of explanations represented as multi-sentence answers for diverse questions where users are encouraged to provide answers, which are comprehensible for a five-year-old. WorldTree V2[2] (Jansen et al., 2019) is a corpus of Science-Domain that contains explanation graphs for elementary science questions, where explanations represent interconnected sets of facts. CoS-E[3] is a dataset of human explanations for commonsense reasoning in the form of natural language sequences and highlighted annotations (Rajani et al., 2019). Multimodal Explanations Datasets (VQA-X and ACT-X) contain textual and visual explanations from human annotators (Park et al., 2018). e-SNLI[4] is a corpus of explanations built on the question: "Why is a pair of sentences in a relation of entailment, neutrality, or contradiction?" (Camburu et al., 2018). Finally, the SNLI corpus[5] is a large annotated corpus for learning natural language inference (Bowman et al., 2015), considered one of the first corpora of NL explanations.

In this paper, we present a new corpus for NL explanations. The ExBAN corpus presented here provides a valuable addition to this set of corpora as it is aimed at explaining structured graphical models (in particular Bayesian Networks), that are closely linked to ML methods.

3 ExBAN Corpus

The ExBAN Corpus (**Ex**planations for **BA**yesian **N**etworks)[6] consists of NL Explanations collected in a two step process: (1) NL explanations were produced by human subjects; (2) in a separate study, these explanations were evaluated in terms of Informativeness and Clarity.

For Step 1, each subject was shown graphical representations of three Bayesian Networks (BN), in random order. They were then asked to produce text to describe how they interpreted the BN. The three BN used in the data collection are presented in Figure 1and represent well-known BN examples, extracted from Russell (2019). For Step 2 in a separate experiment, approximately 80 of these generated explanations were presented to a different set of subjects in random order, along with a scenario description and the graphical model image. Subjects were asked to rate them in terms of Informativeness and Clarity. The worded scenario descriptions were not given to subjects in the first stage, so as not to prime them when generating explanations.

3.1 Step 1: Natural Language Explanations Corpus

Survey Instrument. A pilot was performed to test options and ensure the completion time, leading to the final survey instrument. The survey was divided into five sections: 1) consent form; 2) closed-ended questions related to English proficiency, computing and AI experience: "How much computing experience do you have?", "What is your English Proficiency Level?", "How much experience do you have in the field of Artificial Intelligence?"; 3) attention-check question, where participants received an image of a graphical model, and they had to select the correct answer(s) for the given image; and 4) respondents were asked to explain the three graphical models, in their own words. All respondents received the graphical model survey questions in randomised order. The appropriate ethical procedures were followed in accordance with ethical standards, and ethical approval was obtained.

Participants. 85 participants were recruited via social media. English proficiency level, computing experience and AI experience were rated on a numerical scale, from 1 to 7 (1 = beginner, 7 = advanced). The majority of participants ($n = 83$) rated their level of English proficiency with values higher than 5, with over half of the participants rating their level as 7. Just 12% ($n =$

[1] https://facebookresearch.github.io/ELI5/

[2] http://www.cognitiveai.org/explanationbank

[3] https://github.com/salesforce/cos-e

[4] https://github.com/OanaMariaCamburu/e-SNLI

[5] https://nlp.stanford.edu/projects/snli/

[6] The data is openly released at https://github.com/MirunaClinciu/ExBAN

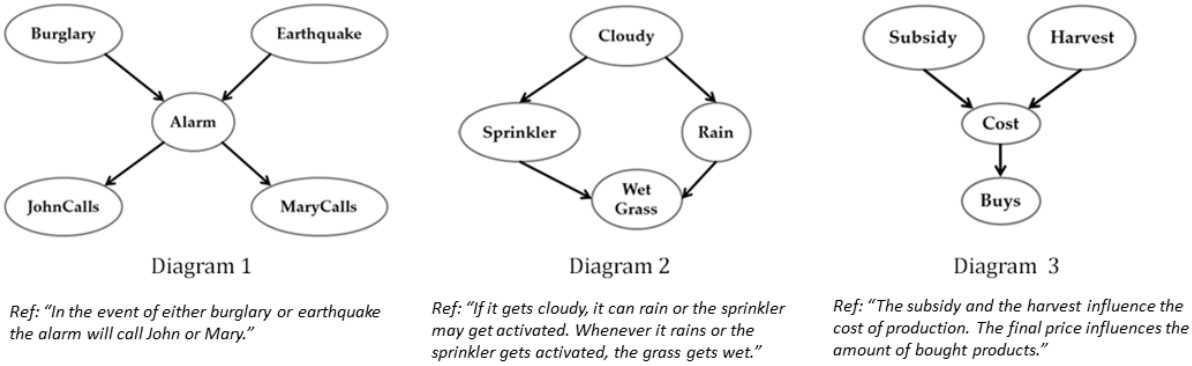

Figure 1. Annotated diagrams with assigned explanation references, where **Diagram 1** represents a typical Bayesian Network, **Diagram 2** represents a multiply-connected network and **Diagram 3** represents a simple network with both discrete variables (Subsidy and Buys) and continuous variables (Harvest and Cost). Beneath each diagram, the gold standard references are provided.

10) participants rated their computing experience scores with a value lower than 5 and 82% ($n = 70$) of participants had a high level of computing experience. Subjects had mixed experience with AI with over half (54%) having some experience ($n = 46$), but 46% of them had limited AI experience ($n = 39$).

Collected NL explanations. Quality control of the collected data included a cleaning step where participants' responses were hand-checked and removed if the participants did not attempt to complete the tasks. Explanations that contained misspellings and missing punctuation were corrected manually (both the raw data and cleaned data are available). The number of explanations for each diagram, after the data cleaning step are as follows: Diagram 1: 84 explanations, 1788 words; Diagram 2: 83 explanations, 1987 words; and Diagram 3: 83 explanations, 1400 words.

3.2 Step 2: Human Evaluation for Quality

Survey Instrument. To investigate the quality of the explanations collected in Step 1, we performed a human evaluation of the generated explanations. A pilot survey was performed to test and refine options, where respondents ($n = 45$) were recruited from Amazon Mechanical Turk and were compensated monetarily.

Each participant was given three tasks, each corresponding to the BN presented in Figure 1 with the order randomised. Along with the BN image, a simple description story was provided in order to give the subject a better understanding of the context as well as instructions on how to approach these tasks. Here, we give the scenario for Diagram 1 to illustrate this: "John and Mary bought their dream home. To keep their home safe, they installed a Burglary/Earthquake Alarm. Also, they received an instruction manual where they found the following diagram: They are not sure if they correctly understood the diagram. On the following pages are some worded explanations. We need your help to evaluate them!"

For every BN image, the participants were asked to evaluate 5 explanations in terms of: **Informativeness** (Q: "How relevant the information of an explanation is"; Likert scale, where 1 = Not Informative and 7 = Very Informative); and **Clarity** (Q: "How clear the meaning of an explanation is"; Likert scale, where 1 = Unclear and 7 = Very Clear).

Participants. The final data collection survey was advertised on social media as "a 10-minute survey, where participants were asked to provide feedback about how understandable the explanations of the three graphs are". Demographic information was collected (age range and gender). A total of 96 participants answered the survey. As screening criteria, participants had to complete all survey questions. Post validation, we had a sample of 56 participants consisting of 42 male participants (75%), 11 female participants (19.6%) and 2 non-binary gender participants (3.6%). Gender imbalance might be due to "differences in female and male values operating in an online environment" (Smith, 2008). Half of the participants ($n = 28$) are aged between 23-29 years old, followed by

30% of participants aged between 18-22 ($n = 17$), 20% aged 40–49 ($n = 11$), 18% aged 30–39 ($n = 10$). Previous studies have identified a high degree of inconsistency in human judgements of natural language (Novikova et al., 2018; Dethlefs et al., 2014); each participant can have a different perception of the interpretation of these metrics, even if a definition of these metrics is provided. Indeed, we found that in our data, explanation ratings can vary significantly, with an explanation rated highly by one person for Clarity, but viewed as very unclear by another annotator. This was the case for both Clarity and Informativeness.

We aim to create a reliable database of varying quality of NL explanations, i.e. where the quality of explanations is generally uncontested by the majority. Therefore, subjective ratings were post-processed. For each explanation, we collected a minimum of 3 judgments. Explanations received ratings from 1 to 7; we classified bad explanations as those with low ratings (ratings <5) and good explanations, as those with higher ratings (ratings ≥ 5). For any one explanation, if the difference between the number of good and bad ratings is ≤ 1, then that explanation is considered hard to judge and difficult to reach a consensus on and thus removed. After this pre-processing step, the corpus contained ratings for 54 explanations for Diagram 1, 34 explanations for Diagram 2, and 54 explanations for Diagram 3.

To verify the agreement between different raters, we used Krippendorff's Alpha, a measure of inter-rater reliability (Krippendorff, 1980). We computed Krippendorff's Alpha coefficient using the Python package `krippendorff` (version 0.3.2). After the pre-processing step, the agreement between subjects increased, see Table 1 for the post-processing Alpha values for each of the Bayes Nets. Alpha values between .21 to .40 indicate fair agreement and values between .41 to .60 indicate moderate agreement (Hallgren, 2012). Here, we can see that explanations for Diagram 2 were particularly contentious, but overall the numbers reflect fair to moderate agreement.

	Diagram 1	Diagram 2	Diagram 3	All Diagrams
Inform.	0.514	0.202	0.420	0.377
Clarity	0.440	0.182	0.361	0.319

Table 1. Inter-annotator agreement measured by Krippendorff's Alpha

4 NLG Evaluation Metrics

Here, we describe the reasoning behind our choice of subjective measures that attempt to capture both the content and its correctness (Informativeness) and quality of expression (Clarity). We also describe objective measures commonly used for automatic evaluation of NLG, and which we will extract from the ExBAN corpus.

4.1 Subjective NLG Evaluation Metrics

Human evaluation is considered a primary evaluation criterion for NLG systems (Gatt and Krahmer, 2018; Mellish and Dale, 1998; Gkatzia and Mahamood, 2015; Hastie and Belz, 2014). Through Explainable AI, we want to achieve Clarity and understanding in communicating the process of AI systems. Therefore, explanations should be clear and easily understood by users. Traditional human evaluation metrics are clearly needed for increasing transparency, avoiding confusion and misunderstanding.

Informativeness. As defined in the field of NLG, Informativeness targets relevance or correctness of an NLG output relative to an input (Dušek et al., 2020). According to the literature, Informativeness can provide "timely, relevant and useful information" (Novikova et al., 2018) and "make information immediately accessible" (Maxwell et al., 2017). Sometimes, Informativeness is linked with accuracy, or adequacy (Novikova et al., 2018). As mentioned previously, explanations contain statements with some prior knowledge that must be accurate and true (Goodrich et al., 2019; Xu et al., 2020).

Clarity. An explanation should be clear to achieve effective communication. In the NLG field, Clarity implies that text is easily understood (Belz and Kow, 2009; van der Lee et al., 2017) and that the reader is familiar with basic information introduced in the text (Lampouras and Androutsopoulos, 2013). In addition, Clarity can also help expose the truthfulness and correctness of textual data (Mahapatra et al., 2016).

4.2 Automatic Evaluation Metrics

This section describes a number of automatic metrics commonly used in NLG evaluation and selected for this study. These fall into two categories: 1) word-overlap metrics, e.g. BLEU, METEOR and ROUGE (Novikova et al., 2017);

and 2) embedding-based metrics, e.g. BERTScore and BLEURT (Sellam et al., 2020). Each of these metrics is compared to one or more "Gold Standard" text as inspired by the Machine Translation community and adopted for evaluating document summarisation and NLG (Belz and Reiter, 2006). The gold standard is normally a piece of natural language text, annotated by humans as correct, i.e. a solution for a given task. Automatic evaluation is based on this gold standard, by verifying potential similarity (Kovář et al., 2016). However, the selection of gold standards involves subjectivity and specificity (Kovář et al., 2016), and this is part of the reason that automatic metrics have received some criticism (Hardcastle and Scott, 2008).

BLEU (B) (Papineni et al., 2001) is widely used in the field of NLG and compares n-grams of a candidate text (e.g. that generated by an algorithm) with the n-grams of a reference text. The number of matches defines the goodness of the candidate text. **SacreBLEU** (SB) (Post, 2018) is a new version of BLEU that calculates scores on the detokenized text. **METEOR** (M) was created to try to address BLEU's weaknesses (Lavie and Agarwal, 2007). METEOR evaluates 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** (R) (Lin, 1971) evaluates n-gram overlap of the generated text (candidate) with a reference. **ROUGE-L** (RL) (Longest Common Subsequence) computes the longest common subsequence (LCS) between a pair of sentences.

BERTScore (BS) (Zhang et al., 2020) is a token-level matching metric with pre-trained contextual embeddings using BERT (Devlin et al., 2019) that matches words in candidate and reference sentences using cosine similarity. **BLEURT** (BRT) (Sellam et al., 2020) is a text generation metric also 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. Evaluation based on the meanings of words using embeddings (BERTScore, BLEURT) might capture some relevant features of explanations, as word representations are dynamically informed by the words around them (McCormick and Ryan, 2019)).

5 Correlation Study of Automatic Metrics

As noted in the introduction, it remains an open question as to what degree the automatic metrics for NLG reviewed above can capture the quality of NL explanations (Clinciu and Hastie, 2019). Thus, we ran a correlation analysis to investigate the degree to which each of the automatic metrics correlates with human judgements using the ExBAN corpus, and which aspects of human evaluation (Clarity/Informativeness), such automatic measures can capture. With regards to the choice of gold standard text, we picked explanations that received the maximum score in the human evaluation, in both Clarity and Informativeness. Gold standard explanations of each diagram are presented in Figure 1.

5.1 Results

The correlations between automatic metrics and human ratings were computed using the Spearman correlation coefficient. For each explanation, we calculated the median of all the ratings given (median was calculated because the data is ordinal, non-parametric rating data, as is also reported in Braun et al. (2018); Novikova et al. (2017)). These medians were then correlated with the automatic metric scores in Tables 2 and 3 and Figure 2. A summary of the results of the correlation analysis include the following:

1. Word-overlap metrics such as BLEU (n = 1,2,3,4), METEOR and ROUGE (n = 1,2) presented low correlation with human ratings.
2. BERTScore and BLEURT outperformed other metrics and produced higher correlation with human ratings than other metrics on all diagrams. BERTScore values range between [0.23, 0.43] and for BLEURT values range between [0.26, 0.53].
3. Human ratings for Informativeness and Clarity are highly correlated with each other, as observed in Figure 2 ($r = 0.82$).

5.2 Discussion

BLEU-based metrics can be easily and quickly computed; however, they do not correlate as well with human ratings as other methods presented here. This might be due to certain limitations, such as the fact that they rely on word overlap

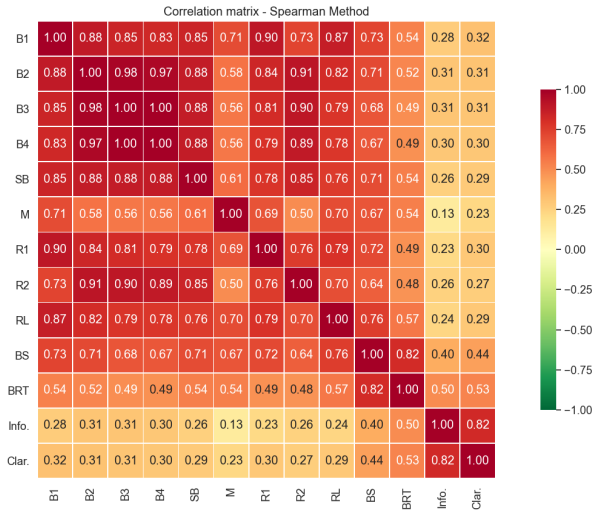

Figure 2. Heatmap of Spearman rank correlation between automatic evaluation metrics and human evaluation metrics (Informativeness and Clarity)

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

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

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

Table 3. Spearman correlation between automatic evaluation metrics and human ratings for Clarity, where the bold font represents the highest correlation coefficient obtained by an automatic evaluation metric

and are not invariant to paraphrases. Furthermore, they do not use recall, rather a Brevity Penalty, which penalizes generated text for being "too short" (Papineni et al., 2001). This way may not be appropriate for explanations, as good explanations may need to be lengthy by their very nature.

METEOR takes into consideration F1-measure by computing scores for unigram precision and recall. The fragmentation penalty is calculated using the total number of matched words (m, averaged over hypothesis and reference) and the number of chunks. In this way, it could identify synonyms, but perhaps not as well as the embedding-based metrics, as evidenced by the correlation figures in our results. With regards to ROUGE-based scores, due to the upper bound issues presented by Schluter (2017), it is impossible to obtain perfect ROUGE-n scores. Furthermore, ROUGE-L cannot differentiate if the reference and the candidate have the same longest common subsequence (LCS), but different word ordering. Again, word ordering may be important for the explanation in terms of explainee scaffolding (Palincsar, 1986).

It has been brought into question whether a single automatic measure is able to capture multiple aspects of subjective human evaluation (Belz et al., 2020). Thus, in order to understand to what degree the various metrics capture both Clarity and Informativeness, we investigated individual explanations and their ratings. Table 4 gives some extracts from the dataset along with the automatic metrics and the human evaluation scores of Informativeness and Clarity. Based on these human scores, the extracts are divided into: good explanations (high scores for both), bad explanations (low scores for both) and mixed explanations (mixed scores). We can see here that all metrics are reasonably good at capturing and evaluating the 'bad' explanations with low scores across the board. However, only the BLEURT metric is good at capturing both 'good and bad' explanation ratings, as observed in the difference in scores between these two categories. ROUGE-L and BERTScore do capture this difference in some cases, but they are not as consistent as BLEURT. The reason that BLEURT outperforms the other metrics may be because it uses a combination of word-overlap metrics as well as embeddings and thus may be capturing the best of these approaches.

Good Explanations	B1	B2	B3	B4	SB	M	R1	R2	RL	BS	BRT	Inf.	Clar.
(1) The alarm is triggered by a burglary or an earthquake.	0.19	0.12	0.00	0.00	0.05	0.23	0.25	0.09	0.12	0.51	0.52	7	7
(2) Cloudy weather may produce rain and activation of the sprinkler. Both rain and sprinkler activity makes the grass wet.	0.28	0.11	0.00	0.00	0.05	0.15	0.36	0.10	0.28	0.49	0.65	7	7
(3) Cost is dictated by the harvest (e.g. size) and available subsidies (e.g. government tax break/subsidy). Whether or not the product is bought depends on the cost.	0.18	0.09	0.00	0.00	0.02	0.07	0.25	0.09	0.12	0.20	0.51	7	7
Bad Explanations	B1	B2	B3	B4	SB	M	R1	R2	RL	BS	BRT	Inf.	Clar.
(4) Sensors = Alarm = prevention or ALERT.	0.06	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.04	0.00	0.00	1	1
(5) A diagram detailing a system whose goal is to make grass wet.	0.08	0.00	0.00	0.00	0.02	0.00	0.11	0.00	0.12	0.13	0.00	1.5	2
(6) The harvest and subsidy contribute to the cost, cost then buys??	0.28	0.00	0.00	0.00	0.03	0.15	0.36	0.00	0.24	0.20	0.30	2	1.5
Mixed Explanations	B1	B2	B3	B4	SB	M	R1	R2	RL	BS	BRT	Inf.	Clar.
(7) The grass is getting wet.	0.08	0.00	0.00	0.00	0.00	0.20	0.13	0.00	0.15	0.24	0.16	1.5	7
(8) Subsidy and harvest independently affect cost. Cost affects buys.	0.06	0.00	0.00	0.00	0.01	0.14	0.16	0.00	0.10	0.25	0.56	6	2.5
(9) Cloud cover influences whether it rains and when the sprinkler is activated. When either the sprinkler is turned on or when it rains, the grass gets wet.	0.48	0.33	0.21	0.14	0.22	0.24	0.50	0.24	0.38	0.49	0.65	7	3

Table 4. Examples of Good, Bad and Mixed Explanations according to human evaluation scores for Informativeness and Clarity (medians of all ratings for that explanation), presented with their automatic measures

Although Clarity and Informativeness highly correlate overall, there are occasions where explanations are rated by humans as higher on Clarity than Informativeness and visa-versa. However, there are rarely any cases where Clarity is high, and Informativeness is very low. Explanation 8 in Table 4 is the only example of this in our corpus. It is thus difficult to make any generalisations about this subset of the data. However, it does seem to be the case that BLEURT is more sensitive to Informativeness than Clarity (e.g. explanation 7 vs 8-9 in the table), but a larger study would be needed to show this empirically.

6 Conclusions and Future work

Human evaluation is an expensive and time-consuming process. On the other hand, automatic evaluation is a cheaper and more efficient method for evaluating NLG systems. However, finding accurate measures is challenging, particularly for explanations. We have discussed word embedding techniques (Mikolov et al., 2013; Kim, 2014; Reimers and Gurevych, 2020), which enable the use of pre-trained models and so reduces the need to collect large amounts of data in our domain of explanations, which is a challenging task. The embedding-based metrics mentioned here perform better than the word-overlap based ones. We speculate that this is in part due to the fact that the former capture semantics more effectively and are thus more invariant to paraphrases. These metrics have also been shown to be useful across multiple tasks (Sellam et al., 2020) but with some variation across datasets (Novikova et al., 2017). Therefore, future work would involve examining the effectiveness of automatic metrics across a wider variety of explanation tasks and datasets, as outlined in the Related Work section.

Embeddings are quite opaque in themselves. Whilst some attempts have been made to visualise them (Li et al., 2016), it remains that embedding-based metrics do not provide much insight into what makes a good/bad explanation. It would thus be necessary to look more deeply into the linguistic phenomena that may indicate the quality of explanations. In ExBAN, initial findings show that the number of nouns and coordinating conjunctions correlate with human judgements, however further in-depth analysis is needed. Additional metrics to add to the set explored here could include grammar-based metrics, such as readability and grammaticality, as in the study described in (Novikova et al., 2017).

Furthermore, an investigation is needed into the pragmatic and cognitive processes underlying explanations, such as argumentation, reasoning, causality, and common sense (Baaj et al., 2019). Investigating whether these can be captured automatically will be highly challenging. We will explore further the idea of adapting explanations to the explainee's knowledge and expertise level, as well as the explainer's goals and intentions. One such goal of the explainer could be to maximise the trustworthiness of the explanation (Ribeiro et al., 2016). How this aspect is consistently subjectively and objectively measured will be an interesting area of investigation.

Finally, the ExBAN corpus and this study will inform the development of NLG algorithms for NL explanations from graphical representations. We will explore NLG techniques for structured data, such as graph neural networks and knowledge graphs (Koncel-Kedziorski et al., 2019). Thus the corpus and metrics discussed here will contribute to a variety of fields linguistics, cognitive science as well as NLG and Explainable AI.

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).

References

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *CoRR*, abs/1909.03012.

Ismaïl Baaj, Jean-Philippe Poli, and Wassila Ouerdane. 2019. Some insights towards a unified semantic representation of explanation for eXplainable artificial intelligence. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 14–19. Association for Computational Linguistics.

Anja Belz and Eric Kow. 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*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Braun, Ehud Reiter, and Advaith Siddharthan. 2018. SaferDrive: An NLG-based behaviour change support system for drivers. *Natural Language Engineering*, 24(4).

Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume Part F147615.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31, pages 9539–9549. Curran Associates, Inc.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108, Tilburg University, The Netherlands. Association for Computational Linguistics.

Miruna-Adriana Clinciu and Helen Hastie. 2019. A Survey of Explainable AI Terminology. *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 8–13.

European Commission. 2018. Article 22 EU GDPR "Automated individual decision-making, including profiling". https://www.privacy-regulation.eu/en/22.htm. Accessed on 2021-01-25.

Lorand Dali, Delia Rusu, Blaž Fortuna, Dunja Mladenić, and Marko Grobelnik. 2009. Question answering based on semantic graphs. In *CEUR Workshop Proceedings*, volume 491.

Maartje M.A. De Graaf and Bertram F. Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *AAAI Fall Symposium - Technical Report*, volume FS-17-01 - FS-17-05.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 702–711. Association for Computational Linguistics (ACL).

Daniel Deutch, Nave Frost, and Amir Gilad. 2016. Nlprov: Natural language provenance. In *Proceedings of the 42nd International Conference on Very Large Data Bases (VLDB) Endowment*, volume 9, page 1537–1540. VLDB Endowment.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Finale Doshi-Velez and B. Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:1–64.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly: Management Information Systems*, 23(4).

Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1).

Maaike Harbers, Karel Van Den Bosch, and John Jules Ch Meyer. 2009. A study into preferred explanations of virtual agent behavior. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5773 LNAI.

David Hardcastle and Donia Scott. 2008. Can we evaluate the quality of generated text? In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for NLG in interactive systems. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Peter A. Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. 2019. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 2732–2740. European Language Resources Association (ELRA).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Vojtěch Kovář, Miloš Jakubíček, and Aleš Horák. 2016. On evaluation of natural language processing tasks: Is gold standard evaluation methodology a good solution? In *ICAART 2016 - Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, volume 2, pages 540–545. SciTePress.

Samantha Krening, Brent Harrison, Karen M. Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. 2017. Learning From Explanations Using Sentiment and Advice in RL. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1).

Klaus Krippendorff. 1980. *Metodología de análisis de contenido. Teoría y práctica.* SAGE, 2004.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 8730–8742, Online. Association for Computational Linguistics.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Gerasimos Lampouras and Ion Androutsopoulos. 2013. Using integer linear programming for content selection, lexicalization, and aggregation to produce compact texts from OWL ontologies. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 51–60, Sofia, Bulgaria. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

David B. Leake. 2014. *Evaluating Explanations*. Psychology Press.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 1971. ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew. *Information Sciences Institute*, 34(12).

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257.

Prashan Madumal, Liz Sonenberg, Tim Miller, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, volume 2.

Joy Mahapatra, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2016. Statistical natural language generation from tabular non-textual data. In *INLG 2016 - 9th International Natural Language Generation Conference, Proceedings of the Conference*.

Steven Mascaro, Ann Nicholson, and Kevin Korb. 2014. Anomaly detection in vessel tracks using Bayesian networks. In *International Journal of Approximate Reasoning*, volume 55.

David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2017. A study of snippet length and informativeness behaviour, performance and user experience. In *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chris McCormick and Nick Ryan. 2019. Bert word embeddings tutorial. Accessed on 2021-01-25.

C. Mellish and R. Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12(4).

Silvia Metelli and Nicholas Heard. 2019. On bayesian new edge prediction and anomaly detection in computer networks. *Annals of Applied Statistics*, 13(4).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv preprint arXiv:1706.07269*.

Tim Miller, Piers Hower, and Liz Sonenberg. 2017. Explainable AI: beware of inmates running the asylum. In *Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI)*, October, page 363.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena" Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Annemarie Sullivan Palincsar. 1986. The role of dialogue in providing scaffolded instruction. *Educational Psychologist*, 21(1-2):73–98.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-jing Zhu, and Yorktown Heights. 2001. IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation. *Science*, 22176:1–10.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8779–8788. IEEE Computer Society.

Gregory Plumb, Denali Molitor, and Ameet Talwalkar. 2018. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, volume 2018-December.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019)*.

Nils Reimers and Iryna Gurevych. 2020. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016.

Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

S. Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group.

Sasan Saqaeeyan, Hamid Haj Seyyed Javadi, and Hossein Amirkhani. 2020. Anomaly detection in smart homes using Bayesian networks. *KSII Transactions on Internet and Information Systems*, 14(4).

Rosario Scalise, Stephanie Rosenthal, and Siddhartha Srinivasa. 2017. Natural language explanations in human-collaborative systems. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 377–378, New York, NY, USA. Association for Computing Machinery.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

William Smith. 2008. Does Gender Influence Online Survey Participation? A Record-Linkage Analysis of University Faculty Online Survey Response Behavior. Accessed on 2021-01-25.

Kacper Sokol and Peter A. Flach. 2019. Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety. In *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zaid Tashman, Christoph Gorder, Sonali Parthasarathy, Mohamad M. Nasr-Azadani, and Rachel Webre. 2020. Anomaly detection system for water networks in northern ethiopia using bayesian inference. *Sustainability (Switzerland)*, 12(7).

Nicole Tourigny and Laurence Capus. 1998. Learning summarization by using similarities. *International Journal of Phytoremediation*, 21.

Adam White and Artur S. d'Avila Garcez. 2019. Measurable counterfactual local explanations for any classifier. *CoRR*, abs/1908.03020.

Yosef Ardhito Winatmoko and Masayu Leylia Khodra. 2013. Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation. *Procedia Technology*, 11.

Bowen Xu, Zhenchang Xing, Xin Xia, and David Lo. 2017. AnswerBot: Automated generation of answer summary to developers' technical questions. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE 2017)*.

Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas. 2020. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081, Online. Association for Computational Linguistics.

Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. 2011. Most relevant explanation in bayesian networks. *The Journal of Artificial Intelligence Research*, 42(1):309–352.

Jeffrey C. Zemla, Steven Sloman, Christos Bechlivanidis, and David A. Lagnado. 2017. Evaluating everyday explanations. *Psychonomic Bulletin & Review*, 24(5):1488–1500.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.