

EvalNLGEval 2020

1st Workshop on Evaluating NLG Evaluation

Proceedings of the Workshop

December 18, 2020

©2020 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-952148-58-3

Preface

The first workshop on Evaluating NLG Evaluation (EvalNLGEval) is taking place virtually as part of the 13th International Conference on Natural Language Generation (INLG 2020).

The aim of the workshop is to offer a platform for discussions on the status and the future of the evaluation of Natural Language Generation (NLG) systems. This is a special time for our field: NLG research has become one of the most popular areas of computational linguistics, the community has expanded and many new tasks and approaches have recently been introduced. However, evaluation of NLG systems remains a bottleneck, as there is no standard methodology for human evaluation nor acceptable automatic metrics, which can hinder reproducibility and comparability of results. The workshop aims to break ground by initiating discussions around these issues.

The workshop invited archival papers and abstracts on NLG evaluation including best practices of human evaluation, qualitative studies, cognitive bias in human evaluations etc. The workshop received twelve submissions. Archival papers were reviewed by three members of the programme committee. Abstracts were accepted by a unanimous decision of the organization committee based on relevance; in case of conflict of interest, abstracts received two reviews. Ten papers and abstracts were accepted and were presented as posters at the workshop. This proceedings volume contains the five archival papers.

The workshop features a keynote speech by Marina Fomicheva and a panel discussion with Yvette Graham, João Sedoc and Marina Fomicheva on the current limits, as well as the future of NLG evaluation. The posters were presented in four poster sessions and the workshop closes with a general discussion on NLG evaluation.

We would like to thank the authors, the program committee members, and the workshop attendees.

Shubham Agarwal
Ondřej Dušek
Sebastian Gehrmann
Dimitra Gkatzia
Ioannis Konstas
Emiel van Miltenburg
Sashank Santhanam
Samira Shaikh

Organizers:

Shubham Agarwal, Heriot-Watt University
Ondřej Dušek, Charles University
Sebastian Gehrmann, Google AI Language
Dimitra Gkatzia, Edinburgh Napier University
Ioannis Konstas, Heriot-Watt University
Emiel van Miltenburg, Tilburg University
Sashank Santhanam, University of North Carolina at Charlotte
Samira Shaikh, University of North Carolina at Charlotte

Program Committee:

José M. Alonso, University of Santiago de Compostela
Miruna A. Clinciu, Heriot-Watt University
Thiago Castro Ferreira, University of São Paulo
Behnam Hedayatnia, Amazon
David M. Howcroft, Heriot-Watt University
Chris van der Lee, Tilburg University
Saad Mahamood, trivago N.V.
Simon Mille, Universitat Pompeu Fabra
Ehud Reiter, University of Aberdeen
Thibault Sellam, Google
Simeng Sun, University of Massachusetts Amherst
Alex Wang, New York University

Invited Speaker:

Marina Fomicheva, University of Sheffield

Panelists:

Marina Fomicheva, University of Sheffield
Yvette Graham, Dublin City University
João Sedoc, New York University

Table of Contents

A proof of concept on triangular test evaluation for Natural Language Generation	1
<i>Javier González Corbelle, José María Alonso Moral and Alberto Bugarín Diz</i>	
"This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation	10
<i>Stephanie Schoch, Diyi Yang and Yangfeng Ji</i>	
Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation	17
<i>Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira and Emiel Kraemer</i>	
NUBIA: NeUral Based Interchangeability Assessor for Text Generation	28
<i>Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh and Mohamed Coulibali</i>	
On the interaction of automatic evaluation and task framing in headline style transfer	38
<i>Lorenzo De Mattei, Michele Cafagna, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim and Albert Gatt</i>	

Workshop Programme

11:00–11:15 Opening

11:15–12:15 Plenary Keynote by Marina Fomicheva

Think Inside the Box: Glass-box Evaluation Methods for Neural MT

12:15–12:50 Break

12:50–13:20 Elevator pitches for all papers

13:20–13:50 Poster session 1

Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing (abstract)

Brian Thompson and Matt Post

Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents (abstract)

Sashank Santhanam and Samira Shaikh

13:50–14:20 Poster session 2

On the interaction of automatic evaluation and task framing in headline style transfer

Lorenzo De Mattei, Michele Cafagna, Huiyuan Lai, Felice Dell’Orletta, Malvina Nissim and Albert Gatt

Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference (abstract)

Ondřej Dušek and Zdeněk Kasner

14:20–15:00 Break

15:00–16:00 Panel discussion with Q&A

Panelists: Marina Fomicheva, Yvette Graham, João Sedoc

16:00–16:30 Poster session 3

Informative Manual Evaluation of Machine Translation Output (abstract)

Maja Popović

NUBIA: NeUral Based Interchangeability Assessor for Text Generation

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh and Mohamed Coulibali

“This is a Problem, Don’t You Agree?” Framing and Bias in Human Evaluation for Natural Language Generation

Stephanie Schoch, Diyi Yang and Yangfeng Ji

16:30–16:50 Break

16:50–17:20 Poster session 4

A proof of concept on triangular test evaluation for Natural Language Generation
Javier González Corbelle, José María Alonso Moral and Alberto Bugarín Diz

Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation
Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira and Emiel Krahmer

Evaluating AMR-to-English NLG Evaluation (abstract)
Emma Manning, Shira Wein and Nathan Schneider

17:20–18:20 General discussion, closing

A proof of concept on triangular test evaluation for Natural Language Generation

Javier Gonzalez-Corbelle, Jose M. Alonso, A. Bugarín
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain

{j.gonzalez.corbelle, josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

Abstract

The evaluation of Natural Language Generation (NLG) systems has recently aroused much interest in the research community, since it should address several challenging aspects, such as readability of the generated texts, adequacy to the user within a particular context but also moment and linguistic quality-related issues (e.g., correctness, coherence, understandability), among others. In this paper, we propose a novel technique for evaluating NLG systems that is inspired on the triangular test used in the field of sensory analysis. This technique allows us to compare two texts generated by different subjects and to i) determine whether statistically significant differences are detected between them when evaluated by humans and ii) quantify to what extent the number of evaluators plays an important role in the sensitivity of the results. As a proof of concept, we apply this evaluation technique in a real use case in the field of meteorology, showing the advantages and disadvantages of our proposal.

1 Introduction

Evaluation can be defined as “the systematic determination of the merit, value and meaning of something or someone based on criteria with reference to a set of rules” (Scriven, 1991). For some authors, the concept of evaluation appeared in the 19th century with the industrialization process in the U.S. (Castro and Benito Martínez, 2014). Later on, a modern scientific discourse emerged in the field of education that would incorporate terms such as learning objectives or educational assessment (Gullickson, 2003). Nowadays, evaluation has been extrapolated to many areas beyond education and consists of the process of obtaining evidence that allows to judge the degree of achievement of previously established objectives. Nevertheless, despite the technological advances in recent years, there are still certain areas in which the

evaluation process must be carried out by humans and not just based in data-driven metrics. In these cases, it is difficult to avoid subjective judgments in the evaluation process.

Evaluation of an NLG system usually requires checking the degree to which it meets the established language requirements, such as the quality of the texts generated, their correction, their interpretability, syntax, formatting or style. The task of evaluating NLG systems presents difficulties mainly because usually these systems do not produce a single correct output and therefore it is hard to define universally accepted metrics for NLG evaluation. When conducting an NLG evaluation with users, there is no general consensus about what to ask (e.g., “How fluent do you think the text is?” or “How natural do you think the text is?”), how many evaluators should participate in the assessment process, or which specific statistical tests should be applied. Moreover, subjectivity can influence the evaluation results and make them be devoid of statistical significance.

Although some authors have advised against the use of statistical significance testing in corpus linguistics (Koplenig, 2017), there have been several proposals for addressing the effect of human subjectivity and statistical significance in human evaluation for several computational linguistics related tasks. In this regard (van der Lee et al., 2019) presents an overview of statistical significance tests that are conducted in human evaluation in NLG. They summarize also a set of best practices grounded in the literature. In addition, (Artstein and Poesio, 2008) describes a survey of methods for measuring agreement among corpus annotators. Moreover, (Amidei et al., 2019) shows the limits of considering Inter-Annotator Agreement as the only criterion for checking evaluation reliability, and proposes correlation coefficients and agreement coefficients to be used together with the

aim of obtaining a better assessment of the data reliability for human evaluation in NLG. In spite of this, and to the best of our knowledge, in the evaluation of NLG systems, so far, there are no established protocols or standards to successfully minimize the effect of human subjectivity and to ensure that results are reported with statistical significance as exist in other areas, such as for example, Sensory Analysis. In this realm there are well established procedures and rules for the human-based measurement of the sensory characteristics of products (Naes et al., 2010; European Sensory Science Society, 2020) that guarantee the validity of the evaluation results and their statistical significance. Sensory Analysis and the computational theory of perceptions have already been applied to automatic reporting (Quirós et al., 2016).

In this paper, we propose a technique for NLG evaluation that is supported by some of the standards applied in Sensory Analysis. It consists of a manual evaluation that allows to obtain a global assessment of the generated texts, instead of assessing a unique characteristic (e.g., fluency or coherence). As a matter of fact, the new technique minimizes the subjectivity inherent to human evaluation. We also present the experimental results obtained when carrying out a proof of concept of this technique by comparing real texts generated by two different people. The objective of this preliminary experimental study is to analyze in practice the advantages and disadvantages of the proposed technique before applying it to the evaluation of an end-to-end NLG system that is currently under development.

The rest of the manuscript is organized as follows. In Section 2 we provide a summary of the state of the art of NLG evaluation techniques. In Section 3 we introduce some preliminary concepts in the field of Sensory Analysis needed to understand the new evaluation technique proposed in Section 4. In Section 5 we present the experimental setting and reported results. Finally, Section 6 concludes with final remarks and possible future work.

2 Background

Evaluation of NLG systems is very different from other areas, because of the number and type of different dimensions to be considered. Accordingly, it constitutes one of the current challenges in research in the NLG field.

In the review by van der Lee et al. (2019), the authors present a review of the challenges of evaluating NLG systems, the pros and cons of different evaluation approaches, and a guide to good practice in conducting NLG evaluation. They emphasize the need to conduct an evaluation with people (whenever possible), in addition to using several independent evaluation criteria. They also recommend that the number of evaluators required be duly justified, as well as their socio-demographic profile, and preferably that the evaluation panel be designed with the widest possible audience in mind. The random and balanced design of the samples to be evaluated, as well as their number and order, should also be justified in order to minimize possible biases and the subjectivity of results. Finally, regarding statistical analysis, it is proposed to distinguish between exploratory studies (more qualitative) and confirmatory studies (more quantitative and supported by results with statistical significance).

In addition, when evaluating an NLG system, it may be necessary to consider aspects related to the final text generated by the whole system or specific aspects of one or several stages of the generation process (e.g., content determination, lexicalization, surface realization, etc.). Due to the great number and diversity of characteristics to be considered when evaluating an NLG system (e.g., readability of the texts generated, coherence, interpretability, etc.), different strategies can be followed (Barros, 2019): extrinsic versus intrinsic evaluation.

Extrinsic evaluation deals with assessing the impact of the system on users or other tasks, focusing on the effects produced by the system (e.g., assessing the decisions made by users based on the system output). Intrinsic evaluation (e.g., evaluating the degree of fluidity of the texts generated) pays attention to the effectiveness of the system itself.

In addition, we need to distinguish between automatic and manual evaluation (Belz and Reiter, 2006). The former is based on metrics that automatically compare the system-generated text with a human-generated text corpus, while the latter requires the participation of humans. It is worth mentioning that a key issue comes into play in human evaluation: how to handle the subjectivity introduced by the evaluators when judging the system-generated text.

It is quite common for intrinsic evaluation to be

carried out using automatic metrics such as BLEU or ROUGE (Reiter and Belz, 2009) and for extrinsic evaluation to be carried out manually. However, in some cases more than one type of evaluation needs to be considered because they are complementary (Gkatzia and Mahamood, 2015).

When it comes to selecting the evaluation technique, there are a multitude of automatic and manual techniques that can be applied, but unfortunately there is not still an adequate cataloguing and characterization of them. Generally, manual evaluations tend to be more costly in both time and money. A clear example of manual evaluation that involved considerably high costs (20 months and 75,000GBP) was the evaluation of the STOP system (Reiter et al., 2003), which automatically generated personalized letters to encourage users to stop smoking. On the other hand, automatic type metrics are usually cheaper and allow for quick results. However, some aspects of a text generated by an NLG system, such as correctness or consistency, are difficult to evaluate by automatic metrics. In these cases, a manual evaluation is most appropriate, where, usually, evaluators are asked to rate or rank several texts (Tintarev et al., 2016). In other cases, a task-based evaluation is also carried out, whereby evaluators must make a decision based on the output provided by the system (Portet et al., 2009; Gkatzia et al., 2017).

Automatic metrics assess the degree of accuracy or objectively score how good the output of a system is with respect to the evaluated issues. However, when it comes to human evaluation, the main problem is the inherent subjectivity of each evaluator. Therefore, the introduction of standards and or protocols for obtaining objective and statistically significant results in the context of human evaluation would be highly appreciated by the NLG community.

3 Preliminaries

Sensory Analysis is a well-established scientific discipline with a wide range of applications (e.g., tasting cheeses, oils, wines, creams, etc.) and standards for human evaluation, which are developed by the International Organization for Standardization (ISO) (AENOR, 2010; International Standardization Organization, 2019).

In this paper we adapt an evaluation technique from Sensory Analysis to NLG. In order to understand our proposal, basic concepts of Sensory

Analysis are introduced below:

- **Product:** material to be evaluated.
- **Sample:** unit of product prepared, presented and evaluated.
- **Difference:** situation in which samples can be distinguished based on their sensory properties.
- **Similarity:** situation in which the perceptible differences between the samples are so small that the products can be considered interchangeable.
- **α -risk:** probability of concluding that a difference exists when it does not. Although this is a probabilistic value ($\alpha \in [0, 1]$), the usual values of α in the field of Sensory Analysis range from 0.2 to 0.001 depending on the sensitivity required by the test. As a rule of thumb, given a statistically significant result, the lower the α -risk value, the greater the evidence of difference.
 - An α -risk from 0.2 to 0.05 indicates slight evidence of difference.
 - An α -risk from 0.05 to 0.01 indicates moderate evidence of difference.
 - An α -risk from 0.01 to 0.001 indicates strong evidence of difference.
 - An α -risk of less than 0.001 indicates very strong evidence of difference.

The usual values of α in the field of sensory analysis and those we will use in this paper are $\{0.001; 0.01; 0.05; 0.1; 0.2\}$ depending on the sensitivity required by the test.

- **β -risk:** probability of concluding that a difference does not exist when it does. Like the α -risk value, this is a probability value, but the usual values of β are $\{0.001; 0.01; 0.05; 0.1; 0.2\}$. The strength of the evidence that there is no difference given a statistically significant result is determined using the same criteria as for α -risk, only in this case “evidence of difference” is replaced by “evidence of similarity”.
- **p_d :** maximum allowable proportion of subjects who perceive a difference. This param-

eter in the field of Sensory Analysis usually takes values among 50%, 40%, 30%, 20% or 10%. A value of p_d less than 25% is considered a low proportion of people perceiving a difference, while values of p_d exceeding 35% represent a high proportion.

- **Sensitivity:** a general term used to summarize test results. Ability to perceive, identify and/or differentiate qualitatively and/or quantitatively one or more stimuli through sense organs. In statistical terms, test sensitivity is defined by the values of α , β and p_d . For example, if low values of α and β (less than 0.01) are taken and the value of p_d is less than 25%, then the sensitivity of the test is high. Conversely, if the values of α , β and p_d are high (e.g., $\alpha = 0.2$; $\beta = 0.1$ and $p_d = 40\%$), then the sensitivity is low.
- **Triad:** Three samples offered to the judge¹ in the triangle test.
- **Triangle test:** A technique that describes a procedure to determine whether there is a discernible sensory difference or similarity between the samples of two products. Judges are given a triad and informed that two of the samples are the same and one is different. Judges should note the sample they believe to be different.

4 The NLG Triangle Test

The evaluation technique proposed in this paper consists of a triangle test taken from the Sensory Analysis research field and adapted to NLG evaluation. Thus, instead of presenting the judges triads of food samples in which two of them are the same and one is different, they will be shown three text samples, two generated by the same subject and a third generated by a different subject. In this way, the judges will have to identify which one of the text samples in the triad has been written by a different subject from the other two. It is worth noting that this technique is applicable regardless how the texts under consideration were generated, either manually by humans or automatically by NLG systems, i.e., no matter if each subject is a human or an NLG system.

¹In the field of Sensory Analysis, evaluators participating in a test are called judges.

4.1 Guidelines

The steps to carry out for the preparation and application of the NLG triangle test are as follows:

1. **Establishing the goal of the test:** to detect difference or to detect similarity. If we want to prove that there is perceptible difference between the texts of two different subjects, we have to apply a triangle test of difference where the null hypothesis is that there is no perceptible difference and we try to demonstrate through the triangle test the alternative hypothesis: there is difference. In the case of wanting to prove that two texts are similar and that there is no perceptible difference between them, the situation would be the opposite: we set a null hypothesis in which the texts of each subject are considered to be significantly different and we try to demonstrate by means of the test the alternative hypothesis: there is no significant difference between the texts and they could be considered interchangeable.
2. **Determining the number of judges required to perform the test.** This number depends on the desired sensitivity of the test, in terms of α -risk, β -risk and p_d (see table 1²). Alternatively, table 1 can be used to look for the combination of values of α , β , and p_d that provides an acceptable sensitivity given the number of judges available in a particular scenario. By its own definition, the value we select for α and β will be more relevant depending on the type of triangle test (difference or similarity). The value of p_d determines the maximum proportion of subjects that we allow to detect a difference. For example, if we performed a triangle test of similarity with a value of p_d of 20%, we would be trying to detect the case for which no more than 20% of the judges detect difference between the texts to be evaluated.
3. **Preparing the test procedure.** Each judge will evaluate a triad of text samples where two of the texts are written by the same subject and the other text is written by a different subject. Therefore, if we tag the texts generated

²For a test of difference, a minimum of 18 judges is recommended, while for a similarity test the minimum recommended is 30, regardless of the sensitivity required by the test (AENOR, 2010; International Standardization Organization, 2019).

α	p_d	β				
		0.2	0.1	0.05	0.01	0.001
0.2	50%	7	12	16	25	36
0.1		12	15	20	30	43
0.05		16	20	23	35	48
0.01		25	30	35	47	62
0.001		36	43	48	62	81
0.2	40%	12	17	25	36	55
0.1		17	25	30	46	67
0.05		23	30	40	57	79
0.01		35	47	56	76	102
0.001		55	68	76	102	130
0.2	30%	20	28	39	64	97
0.1		30	43	54	81	119
0.05		40	53	66	98	136
0.01		62	82	97	131	181
0.001		93	120	138	181	233
0.2	20%	39	64	86	140	212
0.1		62	89	119	178	260
0.05		87	117	147	213	305
0.01		136	176	211	292	397
0.001		207	257	302	396	513
0.2	10%	149	238	325	529	819
0.1		240	348	457	683	1011
0.05		325	447	572	828	1181
0.01		525	680	824	1132	1539
0.001		803	996	1165	1530	1992

Table 1: Number of judges for the NLG triangle test. This table is taken from (AENOR, 2010), and is an adaptation of the original table in (Schlich, 1993).

by the first subject as A and the texts generated by the second subject as B, there are six possible combinations of triads to be shown to the judges:

ABB ABA AAB
BAA BAB BBA

These triad combinations should be randomly distributed in groups of six among the judges, so that the first six judges evaluate the six different triad combinations, the second group of six judges re-evaluate the six possible triad combinations, and so on. In this way, each combination will be evaluated the same number of times if the number of judges is a multiple of six, and if not, the number of evalua-

tions for each combination will be as balanced as possible. For example, if we had 64 judges, there would be four triad combinations to be evaluated eleven times, while two of the combinations would be evaluated only ten times ($11 \cdot 4 + 10 \cdot 2 = 64$). Ideally, each judge should evaluate only one triad, but if we had a limited number of judges, we may make repeated evaluations. Notice that, this is only applicable in case of a test of difference (repeated evaluations are not allowed in case of a test of similarity).

4. **Conducting the test.** The three samples of each triad must be presented at the same time and in the same way for each judge. Each judge is informed that there are two text samples generated by the same subject and one generated by a different subject. He/she may read the text samples as many times as necessary, before selecting one. This is a forced choice test, so even if a judge does not detect any difference between the three samples, he/she is forced to select one sample.

4.2 Data Analysis

As we will detail below, the analysis of the collected data depends on the type of test that was performed. In both cases, the analysis takes into account the number of correct answers, i.e., the number of cases in which judges were able to identify the different sample (i.e., the text written by a different subject) within the triad.

4.2.1 Test of difference

Table 2 provides the minimum number of correct answers needed in a triangle test of difference to determine that there is a discernible difference between the samples. The values in the table are based on a binomial distribution, so a normal approximation to the binomial distribution can be used to calculate the minimum number of correct answers needed given any number of judges. The formula for this calculation, from which the values in the table are extracted, is the following: $x = (n/3) + z\sqrt{2n/9}$, where n is the number of judges in the test, z varies with the level of significance (e.g, $z = 0.84$ for $\alpha = 0.2$; $z = 1.28$ for $\alpha = 0.1$; $z = 1.64$ for $\alpha = 0.05$; $z = 2.33$ for $\alpha = 0.01$; $z = 3,09$ for $\alpha = 0.001$)³ and the mini-

³We considered here the values of z corresponding to the most common values of α or β in Sensory Analysis. How-

imum number of correct answers to determine that there is perceptible difference between the samples is the nearest integer greater than x .

n	α				
	0.2	0.1	0.05	0.01	0.001
6	4	5	5	6	-
7	4	5	5	6	7
8	5	5	6	7	8
9	5	6	6	7	8
10	6	6	7	8	9
11	6	7	7	8	10
12	6	7	8	9	10
13	7	8	8	9	11
14	7	8	9	10	11
15	8	8	9	10	12
16	8	9	9	11	12
17	8	9	10	11	13
18	9	10	10	12	13
19	9	10	11	12	14
20	9	10	11	13	14
21	10	11	12	13	15
22	10	11	12	14	15
23	11	12	12	14	16
24	11	12	13	15	16
...					

Table 2: Minimum number of correct answers needed to conclude that there is perceptible difference. This table is taken from (AENOR, 2010), and is an adaptation of the original table in (Meilgaard et al., 1991).

Optionally, a lower one-sided confidence interval can be calculated for the proportion of the population that can perceive difference between the texts by the following calculation: $1.5 \cdot x/n - 0.5 - 1.5z \cdot \sqrt{(x/n) \cdot (1 - (x/n)) / n}$, where x is the number of correct answers, n is the number of judges, and z varies with the level of significance ($z = 1.28$ for $\alpha = 0.1$; $z = 1.64$ for $\alpha = 0.05$; $z = 2.33$ for $\alpha = 0.01$)³.

4.2.2 Test of similarity

Table 3 shows the maximum number of correct answers allowed to conclude that two samples are

ever, the statistical methods that allow the calculation of z for any other value of α or β are described in more detail by (Meilgaard et al., 1991).

similar for a given number of judges. This table is also based on a binomial distribution, so for any number of judges the upper confidence limit of $100 \cdot (1 - \beta)\%$ can be calculated for p_d using the following normal approximation to the binomial distribution: $1.5 \cdot x/n - 0.5 + 1.5z \cdot \sqrt{(n \cdot x - x^2)/n^3}$, where x is the number of correct answers, n is the number of judges chosen for the test and z varies with the level of significance ($z = 0.84$ for $\beta = 0.2$; $z = 1.28$ for $\beta = 0.1$; $z = 1.64$ for $\beta = 0.05$; $z = 2.33$ for $\beta = 0.01$; $z = 3.09$ for $\beta = 0.001$)³. If the calculated value is below the limit selected for p_d , the samples are declared similar at the β level of significance.

n	β	pd				
		10%	20%	30%	40%	50%
18	0.001	0	1	2	3	5
	0.01	2	3	4	5	6
	0.05	3	4	5	6	8
	0.1	4	5	6	7	8
	0.2	4	6	7	8	9
24	0.001	2	3	4	6	8
	0.01	3	5	6	8	9
	0.05	5	6	8	9	11
	0.1	6	7	9	10	12
	0.2	7	8	10	11	13
30	0.001	3	5	7	9	11
	0.01	5	7	9	11	13
	0.05	7	9	11	13	15
	0.1	8	10	11	14	16
	0.2	9	11	13	15	17
36	0.001	5	7	9	11	14
	0.01	7	9	11	14	16
	0.05	9	11	13	16	18
	0.1	10	12	14	17	19
	0.2	11	13	16	18	21
...						

Table 3: Maximum number of correct answers needed to conclude that two samples are similar. This table is taken from (AENOR, 2010), and is an adaptation of the original table in (Meilgaard et al., 1991).

5 Use Case

With the aim of developing a proof of concept of the technique presented in the previous section, we have applied the NLG triangle test to texts in the meteorological field, an area in which we had designed an NLG system previously (Ramos-Soto et al., 2015). However, it is worth noting that the

texts generated in (Ramos-Soto et al., 2015) are weather forecasts by the local council and do not take into account the whole region. For the sake of simplicity in the recruiting of judges, our use case deals with texts which describe the weather forecast for the entire region and are written by meteorologists.

For illustrative purpose, we considered expert judges (see section 5.1) and non-expert judges (see section 5.2). Judges were asked to fill in a questionnaire (Corbelle, 2020) which is divided into several questions. In each question (see Fig. 1), three meteorological situations were presented. Each situation consisted of an image showing the state of the sky for one day in Galicia and a short text written by a meteorologist from the Meteorological Observation and Prediction Unit of the Galician Meteorological Agency (MeteoGalicia⁴). Of the three situations presented in each question, there were two in which the descriptive text had been created by the same subject and a third in which the creator was a different subject. The judge had to select the text that he/she believed to be created by a different subject from the one who had written the other two.

5.1 NLG triangle test with expert judges

The panel of expert judges was made up of four members of the Non-Linear Physics Group of the University of Santiago de Compostela⁶. The justification for the choice of this group of experts is that they are experts in numerical climate, oceanographic and meteorological models, and therefore very familiar with the vocabulary used in the texts to be evaluated. Moreover, they are independent of the meteorologists who generated the texts to evaluate.

The first step is to determine what type of test (i.e., difference or similarity) should be performed. In our case, due to the small number of judges, we opted for a test of difference in which the repeated evaluations of each judge were considered as if they were independent evaluations.

Secondly, the number of judges required is determined based on the desired sensitivity of the test. Again, the small number of experts available forced us to choose 24 judges (i.e., 6 repeated assessments

⁴<https://www.meteogalicia.gal/>


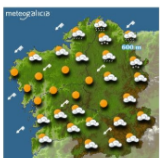

⁵In the questionnaire, the original texts were in Spanish. We provide in the Figure the English translation.

⁶<https://www.usc.gal/en/investigacion/grupos/gfnl/>

Question 1

Select the text you think has been written by a different subject:

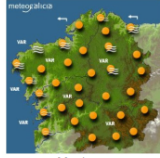
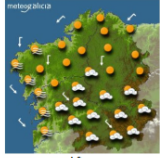

Situation 1

Morning Afternoon Night

Alternating clouds and clear skies in general, with the possibility of occasional rains in the northeast of Lugo, where the snow level will be 600 metres.


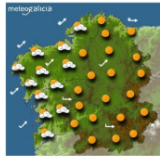
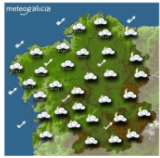
Situation 2

Morning Afternoon Night

In the morning, cloudy or clear skies, with coastal fogs and fog banks in inland areas. In the afternoon, evolution clouds will grow inland.

Situation 3

Morning Afternoon Night

The day will start with low clouds and fog in inland valleys. Both clouds and fog will move backwards to leave a generally very open sky afternoon. The clouds will return at night leaving some light rain in the northwest.

Next

Figure 1: Triangle test questionnaire ⁵

of 4 experts) and then look for a combination of α , β and p_d values that would provide an acceptable sensitivity. In this case, table 1 shows that we can take $\alpha = 0.05$, $\beta = 0.05$, and $p_d = 50\%$ with a minimum number of 23 judges. These sensitivity values assure that the test had a 95% probability ($100 \cdot (1-\beta)$) of detecting the case for which 50% of the judges (i.e., 12 of the 24 judges) can appreciate difference between the test samples.

Accordingly, we can conclude that 50% of the judges could appreciate difference between the samples. It is worth noting that these results did not confirm that there was any similarity between the samples, but simply denied that 50% of the judges were able to perceive difference between the texts.

5.2 NLG triangle test with non-expert judges

In this case we performed a test of similarity. The questionnaire (see Fig. 1) was presented to the general public (non-expert judges) and we had 98 participants. Since repeated evaluations are not allowed in this test, we had 98 evaluations (36 correct), including 16 answers from four of the possible sample combinations and 17 answers from the

remaining two combinations ($16 \cdot 4 + 17 \cdot 2 = 98$). Because of 98 is not a multiple of 6, full balancing of evaluations was not possible.

From Table 3, we can state that having 98 judges, $\alpha = 0.05$, $\beta = 0.01$ and $p_d = 30\%$, the upper confidence limit of $100 \cdot (1 - \beta) = 99\%$ for $p_d = 30\%$ is calculated using the number of correct answers: $1.5 \cdot 36/98 - 0.5 + 1.5 \cdot 2.33 \cdot \sqrt{(98 \cdot 36 - 36^2)/98^3} = 0.221$. Accordingly, we can conclude, with a 99% confidence level, that no more than 22.1% of the judges can detect difference between the compared samples. Therefore, it can be concluded with 99% confidence level that no more than 30% of the population is capable of detecting difference.

6 Final Remarks and Future Work

We have proposed in this paper a new technique for the evaluation of NLG systems. This technique allows us to obtain statistically significant results with the least possible subjectivity from an evaluation carried out by humans, either experts or non-specialists. Our technique provides a mechanism to compare two texts generated by different subjects (either humans or machines) and determines whether difference is detected between them or not.

In the given illustrative use cases, we have learned a number of lessons regarding the type of test. In case of a test of difference, repeated evaluation by judges is allowed. Therefore, each judge can perform several evaluations and be treated as independent. However, in the similarity test repeated assessments are not allowed. Therefore, to obtain equivalent sensitivity levels in a test of difference and in a similarity test, approximately twice as many judges are needed in the similarity test.

We have also seen that and quantified to what extent the number of judges plays an important role in the sensitivity of the results. Although the guidelines in section 4.1 indicate that first the sensitivity values must be determined and then the number of judges, in practice, it is likely that an unlimited number of judges with the required profile will not be available for the evaluation, and therefore sensitivity values will be decided based on the number of judges available. Therefore, if high sensitivity values are required, then a large number of judges must be available. In any case, if the sensitivity level is imposed a priori by the case of study, it will determine the minimum number of judges needed to perform the NLG triangle test.

As the number of evaluators increases, the degree of confidence in the test results also increases. However, if a specific profile of evaluators is required and their availability is low, even with a not very large number of evaluators it is possible to obtain results with a confidence level that in many cases exceeds 90%. In this case, quantitative evidence would support the quality of the texts produced. If a large enough number of evaluators, confidence levels close to 100% can be achieved applying a triangle similarity test. In this case, the conclusion is that empirical evidence shows that a large part of the population does not detect difference between system-generated texts and human-generated texts. Achieving results in this range may require a very large number of evaluators, which in many practical contexts would make the test unfeasible.

As future work, we will apply our NLG triangle test to the comparison between texts generated by NLG systems and texts generated by humans. We will also aim to extend the evaluation technique by including mechanisms for allowing judges to express their motivation for the answers provided and take into account this additional information in the analysis of results.

Acknowledgments

Jose M. Alonso is a *Ramón y Cajal* Researcher (RYC-2016-19802). This research is supported by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, and ED431G2019/04). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

- AENOR. 2010. *Análisis sensorial*, 2 edition. AENOR (Agencia española de Normalización y Certificación).
- J. Amidei, P. Piwek, and A. Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.
- R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.
- C. Barros. 2019. *Proposal of a Hybrid Approach for Natural Language Generation and its Application to Human Language Technologies*. Ph.D. thesis, Department of Software and Computing systems, Universitat d’Alacant.
- A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11 Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy*. The Association for Computer Linguistics.
- L. Castro and J. Benito Martínez. 2014. Following the concept of educational assessment. *Paradigma: Journal of Educational Research*, 20(33):103–115.
- J. González Corbelle. 2020. *D2T validation, Data to Text systems validation questionnaire*. Accessed July 3, 2020.
- European Sensory Science Society. 2020. [link].
- D. Gkatzia, O. Lemon, and V. Rieser. 2017. Data-to-text generation improves decision-making under uncertainty. *IEEE Computational Intelligence Magazine*, 12(3):10–17.
- D. Gkatzia and S. Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), Brighton, UK*, pages 57–60. The Association for Computer Linguistics.
- A. R. Gullickson. 2003. *The student evaluation standards: How to improve evaluations of students*. Corwin Press.
- International Standardization Organization. 2019. Sensory analysis - general guidance for the application of sensory analysis in quality control, ISO 20613:2019.
- A. Kopleinig. 2017. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*.
- C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, and E. Kraemer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- M. C. Meilgaard, B. T. Carr, and G. V. Civille. 1991. *Sensory evaluation techniques*, 2 edition, page 338. CRC press.
- T. Naes, P.B. Brockhoff, and O. Tomic. 2010. *Statistics for sensory and consumer science*. Wiley.
- F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.
- P. Quirós, J.M. Alonso, and D. Pancho. 2016. Descriptive and comparative analysis of human perceptions expressed through fuzzy rating scale-based questionnaires. *International Journal of Computational Intelligence Systems*, pages 450–467.
- A. Ramos-Soto, A. Bugarin, S. Barro, and J. Taboada. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57.
- E. Reiter and A. Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.
- E. Reiter, R. Robertson, and L. M. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.
- P. Schlich. 1993. *Risk Tables for discrimination Tests. Food Quality and Preference*, 4 edition, pages 141–151.
- M. Scriven. 1991. *Evaluation thesaurus*, 4 edition. Sage.
- N. Tintarev, E. Reiter, R. Black, A. Waller, and J. Redington. 2016. Personal storytelling: Using natural language generation for children with complex communication needs, in the wild. . . . *International Journal of Human-Computer Studies*, 92:1–16.

“This is a Problem, Don’t You Agree?” Framing and Bias in Human Evaluation for Natural Language Generation

Stephanie Schoch* Diyi Yang† Yangfeng Ji*

* Department of Computer Science, University of Virginia, Charlottesville, VA 22904

† College of Computing, Georgia Institute of Technology, Atlanta, GA 30332

{sns2gr, yangfeng}@virginia.edu diyi.yang@cc.gatech.edu

Abstract

Despite recent efforts reviewing current human evaluation practices for natural language generation (NLG) research, the lack of reported question wording and potential for framing effects or cognitive biases influencing results has been widely overlooked. In this opinion paper, we detail three possible framing effects and cognitive biases that could be imposed on human evaluation in NLG. Based on this, we make a call for increased transparency for human evaluation in NLG and propose the concept of human evaluation statements. We make several recommendations for design details to report that could potentially influence results, such as question wording, and suggest that reporting pertinent design details can help increase comparability across studies as well as reproducibility of results.

1 Introduction

Human evaluation is widely considered the gold standard for evaluating natural language generation (NLG), in part because existing automatic metrics display low correlations with human judgments (Belz and Reiter, 2006; Liu et al., 2016; Reiter and Belz, 2009; Novikova et al., 2017). As a result, human evaluation is frequently used to demonstrate state-of-the-art results for generative tasks. However, this has the potential to be problematic due to the lack of consistency in how human evaluation is carried out (Gkatzia and Mahamood, 2015; van der Lee et al., 2019). Beyond producing variability in results, this has implications for validity of human evaluation results due to the influence of evaluation design choices. To address this, a number of papers have proposed recommended best practices for different aspects of NLG human evaluation (Amidei et al., 2019; van der Lee et al., 2019). However, overlooked have been the issues of transparency and the potential for question framing effects and other cognitive biases influencing results.

Cognitive biases refer to heuristics that arise in judgment or decision-making (Tversky and Kahneman, 1974). Framing effects (Tversky and Kahneman, 1981) are types of cognitive biases that refer to *how* something is asked as opposed to *what* is asked. In the context of natural language generation research, these effects refer to the wording of questions asked and accompanying task descriptions and instructions, as opposed to what the target quality is that is being assessed.

In this opinion paper, we demonstrate the lack of transparency in NLG human evaluation through empirically demonstrating the extent to which question wording is not included in evaluation design details, finding that only 15.68% of human evaluation studies in papers we surveyed explicitly reported the actual questions asked. We discuss three types of framing and cognitive biases that could influence results in NLG human evaluation: positive and negative framing, demand characteristics and response bias, and anchoring and adjusting. Using concrete examples from studies in human-computer-interaction and psychology and hypothetical examples for NLG, we demonstrate the importance of including question wording when using human evaluation in NLG. Finally, we propose the concept of “human evaluation statements” and suggest a set of design parameters that should be included pertaining to human evaluation study design.

2 Transparency in Human Evaluation

There is currently no standardized approach or consensus for how human evaluation for NLG should be carried out (Gkatzia and Mahamood, 2015; van der Lee et al., 2019). As a result, it is currently very difficult to compare results across different studies due to the variability in evaluation design. Past efforts to address this have included

overviews of evaluation design practices used during a particular time span (Amidei et al., 2018; Gkatzia and Mahamood, 2015; van der Lee et al., 2019) with corresponding recommendations for best practices (van der Lee et al., 2019) and empirical studies or overviews investigating the effects of different question types and scales (Amidei et al., 2019; Novikova et al., 2018). Consistently, these studies have approached variability as a factor impacting the reliability of results.

However, yet to be addressed is the lack of transparency in how studies are designed and reported, which has implications for comparability across studies, as well as replicability and validity of results. While transparency has yet to be addressed in human evaluation, transparency of data, models, and automatic evaluations is a growing topic of concern in the machine learning and natural language processing communities. Bender and Friedman (2018) proposed the usage of “data statements” for mitigating bias and increasing transparency in natural language processing and Gebru et al. (2020) proposed “datasheets for datasets” for increased data transparency and accountability. Transparency in model reporting has also been advocated for. Mitchell et al. (2019) proposed the usage of “model cards” containing model performance characteristics for transparent model reporting. Pertaining to model evaluation, there have been numerous criticisms of task leaderboards (Linzen, 2020; Rogers, 2019) which has led to calls for transparency through reporting of a more informative suite of metrics (Dodge et al., 2019; Ethayarajh and Jurafsky, 2020).

Driving the call for transparency has been the increased attention to issues of reproducibility. Crane (2018) identified a number of controllable environmental settings that are widely unreported in question answering research and demonstrated the impact they have on reproducibility of results, including whether or not a model would be considered state-of-the-art. When we consider the impact of environmental variables (Crane, 2018), computational budget including number of hyperparameter search trials (Dodge et al., 2019), and other factors that can impact results, we can draw comparisons to human evaluation design details that could similarly impact results.

We suggest that the design details of human evaluations can be thought of analogously to model hyperparameters, in that careful tuning can directly

influence results. It is currently an open question as to what parameters in human evaluation could influence results, but without reporting pertinent details, we cannot begin to make comparisons across studies, or reproduce results. For example, van der Lee et al. (2019) suggested their findings pertaining to sample sizes and demographics in a survey of 89 papers using human evaluation for NLG may not reflect reality, since only 55% of the papers reported the number of participants and 18% reported demographics. An additional design parameter that we believe is largely unreported but could have an immense impact on results is that of the actual wording of questions presented to participants. More specifically, if questions are framed in ways that elicit various cognitive biases such as framing effects, response biases, or anchoring and adjustment effects, results may reflect question design rather than model performance.

Empirical Analysis To identify the extent to which question wording is unreported in the details of human evaluation for NLG, we collected a set of 81 NLG papers published in ACL ($n = 33$), EMNLP ($n = 30$), NAACL ($n = 11$), and INLG ($n = 7$) in 2019 and 2020, randomly sampled from all papers containing the keyword “generation” in the title.¹ Of these, 51 (62.96%) included human evaluation as a means to assess model performance. However, only 8 of the 51 studies (15.68%) that included human evaluation reported the actual wording and setup of the questions that were asked, either written out ($n = 4$), included as a figure displaying the prompt ($n = 3$), or both ($n = 1$). Question wording does not only have implications for increasing transparency for the purposes of comparability of results across studies, but has further implications for the validity and reproducibility of results. In the following section, we bring attention to the potential of framing effects and other cognitive biases to impact the results of human evaluation for NLG, and use this to make a case for reporting question wording as part of study design.

3 Framing Effects and Cognitive Biases

Framing (Tversky and Kahneman, 1981) refers to *how* something is asked as opposed to *what* is asked. In human evaluation for NLG, this would be reflected in the question wording or instructions provided to participants. In this section, we detail

¹Data is available at <https://github.com/stephanieschoch/framing-bias-nlg-eval>

three possible framing effects and cognitive biases that could influence the results of human evaluation: positive and negative framing, demand characteristics, and anchoring and adjusting. As question wording is extensively not reported in human evaluation in NLG, rather than providing empirical examples we provide hypothetical examples of the forms these effects could take when question wording is not reported.

3.1 Positive and Negative Framing

Seminal work on the influence of framing in decision-making by [Tversky and Kahneman \(1981\)](#) demonstrated that people are more likely to make choices that are framed positively (in terms of gains) as opposed to negatively (in terms of losses) due to the increased perceived risk associated with choosing potential losses. This effect has been extended and further demonstrated as “loss aversion” in the field of economics ([Levin et al., 2002](#)). In our context, the concept of framing based on positive or negative aspects can be extended and viewed as the framing of questions to induce positive or negative priming effects, in which participants are primed to view a choice as having more positive aspects than another, i.e. as the *better* option. For example, if fluency is the target quality in an NLG evaluation, we can consider it the positive aspect.

We demonstrate the potential for the effects of imposing positive or negative framing and priming on questions in NLG human evaluation with the following example: Suppose a researcher is evaluating sentence A from their generative model against sentence B from a baseline model. The researcher asks participants to respond to the question:

“How much more fluent is sentence A versus sentence B?”

Framing in this manner can prime participants to view sentence A as having more positive aspects, in this case, more fluency, as opposed to neutrally framed questions such as *“How do sentence A and sentence B compare in terms of fluency?”*. Positive or negative framing could therefore have a direct impact on the results of the study, in other words, the results could reflect the framing rather than the actual model performance.

3.2 Demand Characteristics

Demand characteristics are response biases that refer to cues in a study design that may reveal a

researcher’s hypothesis to the participants, resulting in adjusting responses to meet the expectations of the researcher ([Orne, 1962](#)). [Dell et al. \(2012\)](#) demonstrated participant response bias due to interviewer demand characteristics in evaluating human-computer interactive systems. Specifically, when participants knew which artifact was developed by an interviewer, they were consistently more likely to report preference for it, even when it was inferior. For human evaluation in NLG, if questions are framed in a way that cues the evaluators as to which output corresponds to the researcher’s system, it is probable that similar response bias could be elicited. As an example, in the context of NLG, this could take form as follows:

A researcher has developed style transfer model A to generate formal sentences, and is evaluating sentence A from their generative model against sentence B from a baseline model. Unconsciously aware of model A’s artifacts, in this case, as a system that only uses “.” as end punctuation, the researcher states ‘We consider sentences that end with “.” as more formal than sentences that end with “!’” in the task description.

Framing the question in this manner subjects the responses to demand characteristics as the participants are aware of the researcher’s expectations that they will rank sentences ending with “.” as more formal than sentences with alternative end punctuation. Due to the fact that most studies are conducted via crowdsourcing platforms in which annotators receive compensation for responses, this adds an additional incentive to perform in accordance with the researcher’s expectations.

3.3 Anchoring and Adjusting

Anchoring and adjusting is a cognitive bias in which participants anchor their perceptions based on an initial value and adjust subsequent evaluations accordingly ([Tversky and Kahneman, 1974](#)). [Gehlbach and Barge \(2012\)](#) demonstrated anchoring and adjustment effects on attitude-opinion questionnaires in which participants insufficiently adjusted responses on adjacent questionnaire items measuring similar constructs, which affected scale reliability. In the context of human evaluation for NLG, we present the following scenario in which we extend the concept of framing to include framing of task description and instructions displayed alongside questions to elicit advantageous anchoring effects:

A researcher has developed style transfer model A to generate formal sentences. As model design is an iterative process, the researcher has seen model A’s outputs throughout the model design process. When selecting example formal sentences to include in the evaluation task description and instructions displayed to participants, the researcher inadvertently selects sentences that look similar to the types of outputs generated by model A. These examples become an anchor for participants in evaluating sentences generated by model A and model B.

By unintentionally framing the question instructions in a way that introduces an advantageous anchor, the results could reflect the overall framing and bias that is introduced rather than the objective model performance differences.

4 Human Evaluation Design Statements

Throughout the previous sections, we have provided examples demonstrating the potential question framing that could elicit human evaluation results for NLG that are biased in favor of a particular model. While these examples may at first glance seem implausible and only possible in cases of conscious (explicit) researcher bias in favor of a particular model, it is important to take into consideration the potential for researchers to possess unconscious (implicit) bias whether due to underlying expectations for a model’s performance or due to influences of publication bias. During the peer review process reviewers may default to heuristics to simplify the task of review, including rejecting papers where models do not achieve SOTA results (Rogers and Augenstein, 2020). This can implicitly motivate and incentivize researchers to show their model performs best on the gold standard of evaluation for NLG: human evaluation. We use this example to demonstrate the potential for the current lack of evaluation design details, in particular question wording, to leave the door open for results that have been subject to framing effects and bias which threatens the validity of the results.

We draw attention to these effects in an effort to both increase researcher awareness to their own evaluation study design, decrease the potential for questions framed in ways in which results reflect question framing rather than actual model performance, and increase the amount of transparency in human evaluation to aid in study replicability and comparability. We also suggest that the results for

studies which do not include exact question wording should be viewed through a skeptical lens *as though they could contain researcher imparted bias that could significantly impact results*. Further, we use our demonstration of the potential for framing effects and biases in question wording as support for a call for transparency in human evaluation for NLG through the inclusion of study design details, which can aid in the development of more robust human evaluation guidelines.

When guidelines exist that can reduce the complexity and time required to design human evaluation studies, they are used. For the evaluation of paraphrase generation, Li et al. (2018) included the human evaluation guidelines they used as an appendix, which have since been adopted by other studies (Qian et al., 2019). This example shows that guidelines for human evaluation have value: guidelines make life easier and people often adopt those that are available. As such, we make the case for increased transparency in human evaluation with respect to design details that could potentially influence results. In an effort to take preliminary steps towards human evaluation guidelines, we propose the concept of “human evaluation design statements” akin to data statements (Bender and Friedman, 2018; Gebru et al., 2020) or model cards (Mitchell et al., 2019). Determining what should be included on such statements will require additional input, perspectives, and empirical evidence. As a preliminary effort, we provide a list of design parameters that we believe could influence results and should therefore be included when describing human evaluation design setup:

Question Design: Types, Scales, Wording Basic inclusions pertaining to question design are question type and corresponding scales due to the variability that can arise based on these design decisions (Novikova et al., 2018). Further, as we demonstrated in this paper, question wording also has the potential to influence results. Because of the potential for empirical differences due to how questions are framed, it is imperative to report question wording as part of design details, especially in studies where researchers use human evaluation to claim state-of-the-art performance.

Question Presentation: Ordering, Questions per Annotator Ordering effects are influences on results that occur based on the order in which a sequence of questions is presented (Strack, 1992).

As such, reporting question presentation order or balancing increases transparency as well as study comparability and reproducibility. In addition to ordering effects, response fatigue can occur when the quality and integrity of evaluations degrades as participants tire of a task (Lavrakas, 2008). Due to the possibility of response fatigue effects, statistics regarding the number of questions per annotator should be reported to increase design transparency in terms of potential influences on variability in results.

Target Criteria: Definitions It makes intuitive sense that what is actually being measured in human evaluation would influence results, and further that measuring the same or different target criteria in different studies would impact the comparability of the results. However, naming conventions and definitions are inconsistent and may exhibit significant overlap, such as with naturalness, grammaticality, and fluency (Mir et al., 2019; Novikova et al., 2018). As such, what is being measured should be compared across studies based on definition and the resulting participant understanding of the task, rather than simply based on naming convention: studies may measure the same aspect under different names or different aspects under the same name. Studies consistently reporting this detail in human evaluation is also a preliminary step towards agreed upon task definitions.

Annotators: Demographics, Background, Recruitment, Compensation Understanding and reporting the details of the *human* factor in human evaluation is intuitively one of the most important sets of details to include in terms of transparency and potential influence on results. Inclusions involve who annotators are in terms of demographics and background, how they were recruited, and whether or not annotators received fair compensation (Silberman et al., 2018). As an example impact, annotator familiarity with the target language for a task might largely influence judgments towards biases, fluency, or grammatical correctness. The human factor in human evaluation, our annotators, is central to and interacts with every other detail of study design, and is therefore vital to report.

While this list is not comprehensive, we believe these design details could have influences on evaluation results, and as such, are important details to consider and include.

5 Other Considerations

One of the factors that could limit the potential for widespread adoption of human evaluation statements that include human evaluation design details is the page limits imposed for many journal and conference papers. One approach to combat this is to include the details of human evaluation in Supplementary Material that accompanies papers. However, we suggest that many details in human evaluation design are central to understanding the meaningfulness of results, and further suggest that there will need to be community agreed upon guidelines for what details must be included within main papers. We further suggest that a complementary strategy would be the eventual development of comprehensive, agreed upon human evaluation guidelines that could operate similarly to “long-form” data statements (Bender and Friedman, 2018). In this scenario, guidelines could be referenced, summarized briefly, and appended with pertinent additional study details as was proposed with “short-form” data statements (Bender and Friedman, 2018).

6 Conclusion

In this paper, we demonstrate the extent to which including the details of human evaluation is limited in natural language generation. We further demonstrate the need for including design details such as question wording using existing work in psychology and human-computer interaction on framing and cognitive biases, and cite the recent push for transparency with datasets and model details, such as details of hyperparameter tuning, as support for similar efforts to increase transparency in human evaluation. Based on these observations, we propose working towards human evaluation statements and make several suggested inclusions, while noting the future need for additional perspectives and direct empirical support.

References

- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. [Evaluation methodologies in automatic question generation 2013-2018](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. [The use of rating and Likert scales in natural lan-](#)

- guage generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.
- Anja Belz and Ehud Reiter. 2006. [Comparing automatic and human evaluation of NLG systems](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Matt Crane. 2018. [Questionable answers in question answering research: Reproducibility and variability of published results](#). *Transactions of the Association for Computational Linguistics*, 6:241–252.
- Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. ["yours is better!": Participant response bias in hci](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1321–1330, New York, NY, USA. Association for Computing Machinery.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [Show your work: Improved reporting of experimental results](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. *arXiv preprint arXiv:2009.13888*.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. [Datasheets for datasets](#).
- Hunter Gehlbach and Scott Barge. 2012. Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology*, 34(5):417–433.
- 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.
- Paul J Lavrakas. 2008. *Encyclopedia of survey research methods*. Sage Publications.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Kraemer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Irwin P Levin, Judy Schreiber, Marco Lauriola, and Gary J Gaeth. 2002. A tale of two pizzas: Building up from a basic product versus scaling down from a fully-loaded product. *Marketing Letters*, 13(4):335–344.
- Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. [Paraphrase generation with deep reinforcement learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.
- Tal Linzen. 2020. [How can we accelerate progress towards human-like linguistic generalization?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. [Evaluating style transfer for text](#). 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 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- 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.
- Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11):776.
- Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. [Exploring diverse expressions for paraphrase generation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3173–3182, Hong Kong, China. Association for Computational Linguistics.
- Ehud Reiter and Anja Belz. 2009. [An investigation into the validity of some metrics for automatically evaluating natural language generation systems](#). *Computational Linguistics*, 35(4):529–558.
- Anna Rogers. 2019. [How the transformers broke nlp leaderboards](#).
- Anna Rogers and Isabelle Augenstein. 2020. [What can we do to improve peer review in nlp?](#)
- M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41.
- Fritz Strack. 1992. “order effects” in survey research: Activation and information functions of preceding questions. In *Context effects in social and psychological research*, pages 23–34. Springer.
- Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.
- Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Evaluation rules!

On the use of grammars and rule-based systems for NLG evaluation

Emiel van Miltenburg¹, Chris van der Lee¹, Thiago Castro-Ferreira^{1,2}, and Emiel Krahmer¹

¹Tilburg center for Cognition and Communication, Tilburg University

²Federal University of Minas Gerais (UFMG), Brazil

C.W.J.vanMiltenburg@tilburguniversity.edu

Abstract

NLG researchers often use uncontrolled corpora to train and evaluate their systems, using textual similarity metrics, such as BLEU. This position paper argues in favour of two alternative evaluation strategies, using grammars or rule-based systems. These strategies are particularly useful to identify the strengths and weaknesses of different systems. We contrast our proposals with the (extended) WebNLG dataset, which is revealed to have a skewed distribution of predicates. We predict that this distribution affects the quality of the predictions for systems trained on this data. However, this hypothesis can only be thoroughly tested (without any confounds) once we are able to systematically manipulate the skewness of the data, using a rule-based approach.

1 Introduction

Recent years have seen many Natural Language Generation (NLG) researchers move away from rule-based systems, and towards neural end-to-end systems. These systems are typically evaluated using textual similarity metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014), or ROUGE (Lin, 2004), on large corpora of crowd-sourced texts (e.g., the E2E dataset, Novikova et al. 2016; the WebNLG dataset, Gardent et al. 2017; or MS COCO, Lin et al. 2014). This evaluation strategy tells us to what extent the generated texts are similar to the reference data, but it is often difficult to determine exactly what that resemblance buys us. By now it is well-known that BLEU correlates poorly with human ratings (Elliott and Keller, 2014; Kilickaya et al., 2017; Reiter, 2018; Sulem et al., 2018; Mathur et al., 2020), but BLEU by itself also does not tell us anything about the strengths and weaknesses of a particular model, or model architecture. This paper argues that we need alternative

(or at least additional) metrics to provide this kind of insight. We believe that rule-based approaches are well-suited for this task.

1.1 Not just BLEU; also uncontrolled data

BLEU is an easy target; it's a quick-and-dirty solution that ignores paraphrases and different-but-valid perspectives on the input data. But if we only look at the metrics, we miss the elephant in the room: the corpora we use to train NLG systems are the messy result of underspecified elicitation tasks, where annotators receive very little training as to what the outputs should look like (e.g., van Miltenburg 2016; van Miltenburg et al. 2017). Ideally, we should want training data that conforms to a clear set of guidelines. Having clean data is a means to control the quality of the output of an NLG system. By using crowdsourcing, we have ceded that control to the crowd.¹ The problem with crowdsourcing, and particularly with elicitation tasks to create NLG corpora, is that quality control is difficult. And even if we can control the quality of the data, it is very hard to control the diversity of the generated texts.² This makes it harder to study the ability of NLG systems to generalise from the training data to unseen instances. We will argue (in Section 4) that we need a more systematic approach to produce NLG test benches. We believe

¹Although there are also benefits to having a more uncontrolled elicitation task. For example, having fewer constraints means that the resulting data will be more diverse.

²This is not just a problem in NLG. Freitag et al. (2020, and references therein) describe how human translators tend to produce *translationese*: translations that overly rely on the source text, resulting in less natural-sounding texts. This reduces the diversity of the evaluation data for Machine Translation (MT), which has strong effects on the evaluation metrics used in MT (Freitag et al., 2019). The authors go on to show that we can improve the correlation between modern evaluation metrics and human ratings, by improving the reference data (in this case: asking linguists to generate more fluent and diverse translations). But of course, this kind of exercise is expensive and time-consuming.

that a rule-based approach (combined with new or existing NLG data) would again be ideal.

1.2 The downside of end-to-end systems; opportunities for rule-based approaches

There are many good reasons to develop end-to-end systems. For example, Dušek et al. (2020) found that, in the E2E-challenge (Novikova et al., 2017), sequence-to-sequence models “scored higher than other architectures on word-overlap-based metrics and human-rated naturalness.”³ However, given the above, we can also see the move away from rule-based systems as a means to evade responsibility for whatever output our NLG systems produce. After all: the crowd decides what the output should look like. If we don’t explicitly tell our NLG systems what to do (via rules), we should find other ways to control what the output should look like. And what better way to control and evaluate the output... than to use more rules? This paper presents some ways in which rules and rule-based systems can be used to improve current-day NLG research.⁴

2 Evaluation and cognitive capacities

Ideally, evaluation of NLG systems should be tied to the cognitive capacities those systems are claimed to possess (Schlangen, 2019, 2020).⁵ For example, one could evaluate whether a system is able to produce a grammatically correct sentence. Abilities like these can be formalised as a set of rules (cf. Chomsky’s generative program; Chomsky 1965), and we could simply check whether the output of an NLG system conforms to a pre-defined grammar. Xie et al. (2019) do this using Flickinger’s (2000; 2011) English Resource Grammar, which offers broad coverage of the English language. The DELPH-IN catalogue offers

³Another advantage of neural end-to-end systems that is sometimes mentioned is development speed: if you have a training corpus, you can train an end-to-end system fairly quickly. But Reiter (2020) shows that this advantage is probably overstated (if not false). Elsewhere he remarks that ‘effectively impossible to fix undesirable behaviour in a “deep learning” system’ (Reiter, 2016), meaning such a system would have to be re-trained if any changes need to be made to its output. This makes maintenance very time-consuming.

⁴Code for this paper is available at: <https://github.com/evanmiltenburg/EvaluationRules>.

⁵This may not always be the case. Or at least: not directly. For example, consider the question whether a system is *user-friendly* or *pleasant to use*. Some high-level properties are fairly subjective, and may best be evaluated using human ratings. Still one could argue that these properties may be decomposed into a set of different abilities. For example: using the correct register, being able to translate jargon into layman’s terms, generating unambiguous descriptions.

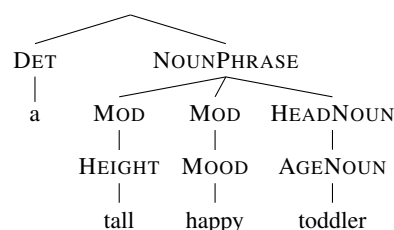


Figure 1: Parse tree for the phrase: *a tall happy toddler*.

an overview of HPSG grammars (Pollard and Sag, 1994) that are available for other languages.⁶ In related work, Bangalore et al. (2000) use automated parsers for evaluation, but they compare the parse trees for the system outputs with those of the reference data, and compute an accuracy metric.

At this point, it is fair to say that not all languages are as well-resourced as English. Of course we can evaluate grammaticality if there is a relatively complete description of a language. But not everyone has that luxury. Moreover, as NLG researchers, we aren’t *just* interested in grammaticality. Why should we care about grammars, then?

There are two ways to respond to this criticism. First, if you accept that NLG evaluation is a good use case for a broad-coverage parser, then that provides additional motivation to build a new or better parser (or to start talking to linguists in your area). Second, a grammar does not necessarily have to cover the *entire* language for it to be useful. It just needs to cover your domain of interest. One example grammar is provided by Van Miltenburg et al. (2018), who developed a context-free grammar to cover person-descriptions in the MS COCO dataset (a collection of images paired with image descriptions). The grammar has a set of production rules that describe as well as categorise the components of person-descriptions in the corpus. So the phrase ‘the tall happy toddler’ can be parsed as in Figure 1. Van Miltenburg (2020) improved this grammar, and used it to evaluate the extent to which image captioning systems are able to generate different kinds of person-descriptions. This is another example of a cognitive ability that can be examined using a pre-determined set of rules. As an additional bonus, a complete characterisation of a domain such as PERSON-DESCRIPTIONS allows us to reflect on what kinds of outputs are desirable or not, for the system to generate.

⁶<http://moin.delph-in.net/GrammarCatalogue>

3 Systems, models, and architectures

Before we continue, it is important to recognise the difference between systems, models and architectures. We consider **architectures** to be abstract descriptions of all the components that make up a system. A **system** is a specific instance that implements an architecture. When a system is trained on a particular dataset, we can say that it has constructed a **model** for how the task should be carried out. These distinctions are important, because different NLG researchers may be interested in either systems, models, or architectures. Theoretically oriented researchers may be more interested in the properties of different architectures, whereas more applied researchers may be interested in the properties of different systems or models. In our experience, shared tasks are often misconstrued as a competition to see who can deliver the best model. This misses the point, because ideally the results of a shared task teach us about the strengths and weaknesses of different architectures.⁷

4 Evaluating the ability to generalise

Machine learning datasets are used to determine whether systems are able to generalise from experience to unseen situations (Mitchell, 1997). To test this, researchers typically use separate training, development, and test sets. Different models are trained using the training set, the best model is selected using the development set, and then we evaluate its performance on the test set.

4.1 Requirements to measure generalisation

Using different splits is necessary, but not sufficient for NLG tasks. We can see this when we look at the generation of weather forecasts, a popular topic in the NLG community (e.g. Gkatzia et al. 2016 and references therein). It is not good enough to only have a corpus where all inputs have the same weather but different place names. NLG models trained on such a corpus would only learn to produce a fixed weather template, where they should copy in the name from the input. An evaluation is only meaningful if there are clear differences in *all* (combinations of) variables, between training and test set. At the same time, the training data should also not contain so much variation that it's impossible to detect any pattern. It is an open question

⁷For further discussion of shared tasks and leaderboards in NLP, see: Parra Escartín et al. 2017; Nissim et al. 2017; Rogers 2019; Ethayarajh and Jurafsky 2020.

how much systematicity (and redundancy) there should be in the training data for NLG systems to learn how to perform any language generation task. Finally, it is important to have specific information about the output. For example: how many different ways are there to verbalise the same predicates, entities, numbers, dates and times? Without this information, it is impossible to say anything about the complexity of the task.

4.2 Are current datasets sufficient?

We don't believe current datasets are sufficient to measure the extent to which systems are able to generalise from the training data, although some datasets do come close. WebNLG, for example, is a state-of-the-art dataset. It offers an excellent overview table (Table 1 in Gardent et al. 2017) describing properties of the input (e.g. number of different predicates, number of combinations of RDF triples, relations between the different triples) and output (e.g. number of sentences verbalising different amounts of triples).

Still missing from the description of the WebNLG corpus is the distribution of different predicates. Figure 2 shows the frequency distribution of different labels in the training set (computed using the XML files from the enriched WebNLG dataset; Castro Ferreira et al. 2018). The plot reveals that the data is heavily skewed, with 76 predicates (out of 246) occurring fewer than 10 times, while the most frequent predicate ('country') occurs 2150 times. End-to-end systems will probably perform worse on the tail of the distribution (where example outputs are scarce) than on the head (where examples are plentiful).

On the output side, it is not clear from the original WebNLG corpus how many different possible lexicalisations there are for each predicate.⁸ This is difficult to study with unstructured text output, but luckily the enriched WebNLG dataset converted the outputs into templates (see Table 2 below), which we can count. Table 1 shows a selection of predicates with different ratios of unique-to-total number of templates. One can imagine that it's much easier for a model to predict the template for a predicate with a ratio of 0.12, than for a predicate with a ratio of 1.00. After all: a lower ratio means that there are more examples for each unique template. The easiest situation would be one where there is a

⁸We limit ourselves to predicates here, but note that predicates are not the only part of the input that needs to be lexicalised.

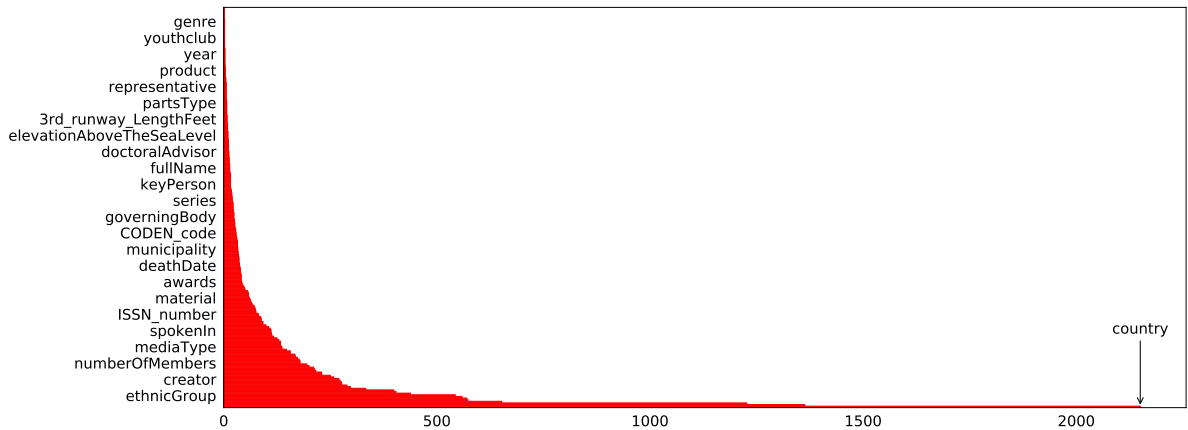


Figure 2: Frequency of all predicates in the training split of the WebNLG corpus. The x-axis shows the frequency of the labels (range: 1–2150, where ‘country’ is the most frequent label). To improve readability, the y-axis only shows a selection of different labels at fixed intervals.

Predicate	Unique	Total	Ratio
fullName	3	3	1.00
product	3	3	1.00
architecturalStyle	17	18	0.94
order	15	16	0.94
discipline	8	9	0.89
champions	18	21	0.86
locationCity	7	9	0.78
demonym	25	32	0.78
foundedBy	2	3	0.67
birthName	6	9	0.67
leaderTitle	37	63	0.59
address	4	7	0.57
areaCode	17	36	0.47
countySeat	7	15	0.47
language	47	119	0.39
city	14	36	0.39
status	3	11	0.27
affiliation	3	11	0.27
capacity	5	26	0.19
capital	11	61	0.18
location	22	177	0.12

Table 1: Number of templates for predicates in the ‘1-triples’ subset of the WebNLG corpus. Columns show the predicate, number of unique templates, total number of templates, and the ratio of unique-to-total number of predicates. This table shows a selection across the entire range of different ratios.

single template, with many examples. Evaluation of NLG systems trained on this corpus should ideally take this uniqueness ratio into account (e.g. by computing performance for different subsets of the data, with different uniqueness ratios).

4.3 What do we need?

Our discussion so far points us in the direction of more carefully planned corpora, with clear input distributions. Ideally the output language should also be controlled, so that it conforms to a set of guidelines for what appropriate output should look like. To really test the degree to which system performance depends on these variables (i.e. the distributions of input and output), having just one big training corpus isn’t good enough. Rather, there should be different versions of the same corpus, so that we can manipulate different aspects of the training data, to see how each of those variables affects the outcome (performance on the test set).

5 How can we get there? Rules!

Corpora constructed solely through human labor are not good enough, because they do not give us enough control over the data to carry out systematic experiments. For example: we only know the amount of different lexicalisations for a predicate *after* data collection has finished. We present two alternative approaches to generate evaluation data.

5.1 From systems to synthetic datasets

One way to construct a controlled corpus, is to use existing NLG systems to produce a large collection of texts in a particular domain. We can then train

end-to-end systems on this data to produce similar texts. This has three major advantages, compared to the use of crowd-sourced data:

1. It is efficient and cost-effective to produce large corpora, since no human annotators are needed.
2. We can easily create different sub-corpora with very specific distributions of the input data, which would allow us to estimate the extent to which systems are able to generalise from low-frequent training examples.
3. It allows us to automatically evaluate the quality of the output in ways that are not possible (or very labor-intensive) with human-generated data.

The generate-and-train approach has recently been applied by Oraby et al. (2018, using the PERSONAGE system; Mairesse and Walker 2010) to create a synthetic corpus of utterances in the RESTAURANT domain, where the authors controlled the *personality* of the utterances. Oraby et al. showed that it is possible for neural NLG models to distinguish style and content, and that models trained on their data were able to generate meaningful output with the desired personality traits.

The idea of training NLG-systems based on the output of other NLG systems is controversial. Ehud Reiter argues on his blog that this is just reverse-engineering existing systems (Reiter, 2017). This is a valid concern if the goal is to build an NLG system to be used in some application context. However, we are not concerned with any applications. Rather, we are interested in the core properties of different end-to-end architectures, and particularly the way those properties relate to learnability: to what extent a particular architecture can learn to generate natural language, based on a corpus with particular, controlled properties?

Feasibility

A natural question at this point is how to find existing systems to generate synthetic NLG corpora. One answer is simply to look for systems using SimpleNLG, since this is probably the most used realisation engine for NLG in academia. It may be possible to build a corpus generation tool that incorporates all different systems. To find these systems and assess the feasibility of our proposal, we used the *Publish or Perish* software⁹ to retrieve all pub-

⁹Search carried out on the 17th of August, 2020, using the macOS GUI edition, version 7.25.2877.7516. Software

Triple: ⟨SAGE_Publications, founder, Sara_Miller_McCune⟩
Text: Sara Miller McCune founded SAGE Publications.
Template: ENT-1 founded ENT-2.
Mapping: ENT-1: Sara Miller McCune
 ENT-2: SAGE Publications

Table 2: Example from the extended WebNLG corpus.

lications on Google Scholar that cite the original SimpleNLG paper (Gatt and Reiter, 2009).¹⁰ We found 361 publications referring to SimpleNLG on Google Scholar, coming from a wide array of different venues. We are still in the process of analysing the results, but our impression is that only a small proportion of the reported systems is useful. Many are either unavailable, form part of a larger pipeline, or use proprietary/personal data (e.g. BT-Nurse; Hunter et al. 2012).

5.2 From datasets to rules, and back again

Another approach is to construct our own template-driven corpus generator, based on existing datasets. Table 2 shows part of an entry from the extended WebNLG corpus. The triple was expressed by a participant through the text ‘Sara Miller McCune founded SAGE Publications.’ Castro Ferreira et al. (2018) semi-automatically converted these texts to templates. Additionally, the dataset also shows how different entities can be realised. This gives us all the ingredients to develop a rule-based system that can generate a corpus matching specific criteria (or indeed a collection of corpora that allow us to determine the ability to which end-to-end systems are able to generalise).

With the templates and entity realisation options in hand, we can choose to make full use of all possible templates and realisations for all predicates and entities, or we can select only specific templates/realisations to have a particular distribution of the data. Here are the aspects that we imagine may be interesting to manipulate:

- The number of different templates/entities in the train, validation, and test sets. (Note that templates and entities may be manipulated separately from each other.)

available from: <https://harzing.com/resources/publish-or-perish>

¹⁰This approach excludes many systems using SimpleNLG in a different language, e.g. Brazilian Portuguese (de Oliveira and Sripada, 2014), Dutch (de Jong, 2018), German (Bollmann, 2011; Braun et al., 2019), French (Vaudry and Lapalme, 2013), Galician (Cascallar-Fuentes et al., 2018), Italian (Mazzei et al., 2016), or Spanish (Ramos-Soto et al., 2017).

- The frequency with which those different templates/entities each occur. Is there a uniform distribution, or do some templates/entities occur more than others?
- The overlap in terms of templates/entities between the train, validation, and test sets. Here we may also choose to generate multiple different test sets to accompany the same training set, to make evaluation more efficient.
- The ordering principles, and the number of different orders in which triples are realised in the output texts. (E.g. maintaining the input order, ordering triples alphabetically or based on their content.)
- The segmentation principles, and the number of sentences that are used to realise a set of triples. (E.g. three triples per sentence; only one triple per sentence; segmentation depending on the predicate, the entities, or both.)
- The amount of noise in the dataset. By default, there is no noise in the synthetic dataset, but we could add synthetic noise (i.e. knowingly introduce errors), to see how systems deal with the presence of noise in the data. This is similar to [Dušek et al. \(2019\)](#), who systematically *removed* noise from the E2E dataset, to gauge the impact of erroneous meaning representations.

Using data generated in this manner, we could answer questions like the following:

- How do skewness and diversity (of templates, referring expressions) influence the quality and diversity of the generated outputs?
- How many minority examples are necessary before end-to-end models consider these a valid alternative to majority examples?
- What kinds of generation rules are learnable by end-to-end systems? Which architectures are more apt to pick up on different kinds of systematic patterns in the data?

Feasibility

There are two main challenges for this approach. The first challenge concerns **multiple predicates**. It is easy to see how textual output for single-predicate inputs can automatically be generated (just fill in the empty slots in the template), but for inputs with multiple predicates the problem is

more complex. The realisation of multiple predicates is not necessarily equal to the realisation of two single predicates, plus some text to link the two (e.g. the conjunction *and*). Indeed, [Perez-Beltrachini et al. \(2016\)](#) purposefully selected combinations of predicates that might lead to more concise solutions. E.g. combining $\langle \text{Alan_Bean, occupation, test_pilot} \rangle$ with $\langle \text{Alan_Bean, nationality, USA} \rangle$ leads to the insertion of an adjective: *Alan Bean was an **American** test pilot.*

Normally it would mean a large amount of human labor to find any systematicity in the corpus. To build a good NLG system, we need to know how to order the predicates; how to relate the predicates to each other; how to split up the information in different sentences; and how to realise sentences combining multiple predicates. However, for evaluation purposes, the exact answers to these questions aren't necessarily important.¹¹ What matters is that there is some output that conforms to a particular set of rules. The evaluation is just there to see if end-to-end systems are able to learn those rules. The exception here is when the ability to learn a specific kind of rule is in question. For example: can neural NLG systems learn to insert adjectives like *American* in the example above?

The second challenge concerns **the distribution of the original corpus**. As Table 1 shows, some predicates occur only three times. This limits the different kinds of corpora that we are able to produce. For example, it is not possible with just the WebNLG data alone to generate a corpus where there are more than three different lexicalisations for the FULLNAME predicate. Moreover, it is not even possible to generate more than nine different predicate-entity combinations (3 entities times 3 predicate-realizations). One way to address the distribution issue is to (semi-)automatically generate more examples by extracting triples from DBpedia ([Auer et al., 2007](#); [Lehmann et al., 2015](#)), and verbalising them using the predicate's lexicalization templates available in the enriched WebNLG dataset and a grammar-based NLG system (e.g., [Mille et al. 2019](#)). As mentioned earlier, the data does not need to be perfect; it just needs to be consistent, so that learners are (in theory) able to infer rules from the data.

¹¹One might have multiple rules corresponding to different answers to each question. It would then be possible to experiment with different amounts of examples generated using the different rules.

5.3 General feasibility

Ideally we would be able to control the data in such a way, that changes to individual variables happen *ceteris paribus*; i.e. with all other variables staying the same. But there are practical considerations we need to take into account:

- The number of times you can train a model, is limited by the size and complexity of that model. If it takes a long time to train the model, then it is not feasible to do this for tens or hundreds of different versions of the same training data.¹²
- This issue is further compounded by the fact that many models are randomly initialised. For a good estimation of system performance, the system needs to be trained multiple times on the exact same data.

There is no universal solution to this problem, but it does help to have specific hypotheses about which factors might affect system performance, and to focus on those.

5.4 Assessing model performance

Next to increased control over the training data, the approaches proposed in this section have an additional benefit: because all training data has been generated using a rule-based approach, we can use those same rules to evaluate which rules were learned by the system, and which ones weren't. This is also the approach we described in Section 2. We could even split up the evaluation, to measure which templates, entity realisations, ordering rules, and segmentation rules the system acquired.

One aspect we did not address yet is how to parse imperfect outputs. There are no guarantees that the output of end-to-end systems will conform to any of the rules through which the corpus was generated. Using a strict approach, we could say that faulty output just doesn't count; if it is flawed, the system simply did not fully learn the relevant rules. But perhaps we would also like to give partial credit to systems that *almost* learned how to perform the generation task. We leave this as a question for future research.¹³

¹²So next to environmental issues caused by computationally heavy approaches to NLP (Strubell et al., 2019), we can also say that such approaches are an obstacle to properly evaluate new systems.

¹³But note that the texts (and probably system outputs as well) are very predictable. This makes it interesting to explore whether metrics based on edit distance could work here, even though they have been shown to be inadequate 'in the wild.'

5.5 Predictions

Since we intend to explore this approach in the future, and to encourage others to explore this space as well, we make a number of predictions:

1. Templates with a lower unique-to-total ratio (see Table 1) are easier to learn.
2. The number of examples needed to successfully learn a template, depends on the amount of alternative templates that could also verbalise the same predicate, the amount of predicates in the corpus, and the size of the corpus.
3. It is easier to learn how to realise a predicate, if the arguments for that predicate are diverse. (If a predicate always occurs with the same arguments, they may be considered part of the template by the model.)
4. When combining multiple triples, conjunction (*Susan is an astronaut and she is American*) is easier to learn than insertion (*Susan is an American astronaut*).

This list is not exhaustive; certainly many more predictions could be made about different combinations of the parameters we described above. But these hypotheses should serve as a starting point for future research. Initially we may want to see empirically whether the predictions hold up for popular architectures. A different avenue of research could be based on this evidence, to develop formal proofs about the properties of families of NLG architectures. We believe both are needed to inform NLG research and practice.

6 Limitations

The output of rule-based systems is often said to be less fluent or natural than the output of end-to-end systems, and this claim is corroborated by the results of the E2E-challenge (Dušek et al., 2020). It may thus be expected that any synthetically generated corpus will be less natural than human-produced data, and the texts will probably have other shortcomings, too. However, the proposal in this paper is focused on determining what systems can or cannot learn from corpora with different properties. This means that, to some extent, the naturalness or fluency of the synthetic data does not really matter. What matters is that we learn how those different properties of the data affect the output of data-driven systems. We can then use those systems in other areas, knowing what they are capable of and what their limitations are. At

that point, we need a different kind of evaluation (although rules are still valuable to check whether system output conforms to particular guidelines).

One might reasonably object here that the quality of the corpus *does* matter. How can you be sure that your synthetic data has the desired properties? Wouldn't that require some form of evaluation as well? We believe this concern could be addressed through unit-tests in the corpus generation code base. Because our proposal involves rule-based generation, the output should always be predictable.

7 Conclusion

We discussed the merits of (grammar) rules and rule-based systems in the context of NLG evaluation. Our conclusion is that there are clear benefits for practitioners who want to learn more about the architectures that they use for real-life applications. A concern that some may have, is that the real world is messy. Why should we solve toy problems like reverse-engineering rule-based systems? Our answer is two-fold. First, since our proposals involve synthetic data, we can make the data as clean or messy as we want. But because we have full control over the data, the evaluation will be much more informative about what systems can or cannot do. Second, a rule-based perspective is useful because it forces us to engage with the data. Looking at the WebNLG data, and all of the different templates that exist for each of the different predicates, one cannot help but ask: is this diversity really useful? Or should we try to reduce the diversity (e.g. formulating guidelines), to ensure the best possible outputs for our NLG systems? Messiness can be good or bad, and it is up to us to explore the impact of variation in NLG data.

Acknowledgments

We thank the anonymous reviewers for their feedback, which helped us refine the arguments laid out in this paper.

References

- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. **Evaluation metrics for generation**. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 1–8, Mitzpe Ramon, Israel. Association for Computational Linguistics.
- Marcel Bollmann. 2011. **Adapting SimpleNLG to German**. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138, Nancy, France. Association for Computational Linguistics.
- Daniel Braun, Kira Klimt, Daniela Schneider, and Florian Matthes. 2019. **SimpleNLG-DE: Adapting SimpleNLG 4 to German**. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 415–420, Tokyo, Japan. Association for Computational Linguistics.
- Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín Diz. 2018. **Adapting SimpleNLG to Galician language**. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 67–72, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Thiago Castro Ferreira, Diego Moussallem, Emiel Kraemer, and Sander Wubben. 2018. **Enriching the WebNLG corpus**. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.
- Michael Denkowski and Alon Lavie. 2014. **Meteor universal: Language specific translation evaluation for any target language**. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. **Semantic noise matters for neural natural language generation**. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.
- 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.
- Desmond Elliott and Frank Keller. 2014. **Comparing automatic evaluation measures for image description**. In *Proceedings of the 52nd Annual Meeting of*

- the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.
- Kawin Ethayarajh and Dan Jurafsky. 2020. [Utility is in the eye of the user: A critique of nlp leaderboards.](#)
- Dan Flickinger. 2000. [On building a more efficient grammar by exploiting types.](#) *Nat. Lang. Eng.*, 6(1):15–28.
- Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In E. M. Bender and J. E. Arnold, editors, *Language from a cognitive perspective: Grammar, usage, and processing*, pages 31–50. Stanford: CSLI Publications.
- Markus Freitag, Isaac Caswell, and Scott Roy. 2019. [APE at scale and its implications on MT evaluation biases.](#) In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.
- Markus Freitag, David Grangier, and Isaac Caswell. 2020. [Bleu might be guilty but references are not innocent.](#)
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data.](#) In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Albert Gatt and Ehud Reiter. 2009. [SimpleNLG: A realisation engine for practical applications.](#) In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece. Association for Computational Linguistics.
- Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. [Natural language generation enhances human decision-making with uncertain information.](#) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–268, Berlin, Germany. Association for Computational Linguistics.
- James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. [Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse.](#) *Artificial Intelligence in Medicine*, 56(3):157 – 172.
- R.F. de Jong. 2018. [Simplenlg-nl : Natural language generation for dutch.](#)
- Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. [Re-evaluating automatic metrics for image captioning.](#) In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleeef, Sören Auer, et al. 2015. [Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia.](#) *Semantic web*, 6(2):167–195.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries.](#) In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft coco: Common objects in context.](#) In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- François Mairesse and Marilyn A. Walker. 2010. [Towards personality-based user adaptation: psychologically informed stylistic language generation.](#) *User Modeling and User-Adapted Interaction*, 20(3):227–278.
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. [Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.
- Alessandro Mazzei, Cristina Battaglini, and Cristina Bosco. 2016. [SimpleNLG-IT: adapting SimpleNLG to Italian.](#) In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK. Association for Computational Linguistics.
- Simon Mille, Stamatia Dasiopoulou, and Leo Wanner. 2019. [A portable grammar-based nlg system for verbalization of structured data.](#) In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC ’19*, page 1054–1056, New York, NY, USA. Association for Computing Machinery.
- Emiel van Miltenburg. 2016. [Stereotyping and bias in the flickr30k dataset.](#) In *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4.
- Emiel van Miltenburg. 2020. [How do image description systems describe people? a targeted assessment of system competence in the people domain.](#) In *Proceedings of LANTERN*. Association for Computational Linguistics.
- Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. [Cross-linguistic differences and similarities in image descriptions.](#) In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

- Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. [Talking about other people: an endless range of possibilities](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 415–420, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Tom M. Mitchell. 1997. *Machine learning*. McGraw-Hill.
- Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. [Last words: Sharing is caring: The future of shared tasks](#). *Computational Linguistics*, 43(4):897–904.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. [The E2E dataset: New challenges for end-to-end generation](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.
- Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. [Crowd-sourcing NLG data: Pictures elicit better data](#). In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK. Association for Computational Linguistics.
- Rodrigo de Oliveira and Somayajulu Sripada. 2014. [Adapting SimpleNLG for Brazilian Portuguese realisation](#). In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 93–94, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.
- Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. [Controlling personality-based stylistic variation with neural natural language generators](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190, Melbourne, Australia. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. [Ethical considerations in NLP shared tasks](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.
- Laura Perez-Beltrachini, Rania Sayed, and Claire Gargent. 2016. [Building RDF content for data-to-text generation](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1493–1502, Osaka, Japan. The COLING 2016 Organizing Committee.
- Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Head-driven Phrase Structure Grammar. University of Chicago Press.
- Alejandro Ramos-Soto, Julio Janeiro-Gallardo, and Alberto Bugarín Diz. 2017. [Adapting SimpleNLG to Spanish](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 144–148, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Ehud Reiter. 2016. Natural language generation and machine learning. Ehud Reiter’s blog, <https://ehudreiter.com/2016/12/12/nlg-and-ml/>.
- Ehud Reiter. 2017. You need to understand your corpora! The Weathergov example. Ehud Reiter’s blog, <https://ehudreiter.com/2017/05/09/weathergov/>.
- Ehud Reiter. 2018. [A structured review of the validity of BLEU](#). *Computational Linguistics*, 44(3):393–401.
- Ehud Reiter. 2020. Is building neural nlg faster than rules nlg? no one knows, but i suspect not. Ehud Reiter’s blog, <https://ehudreiter.com/2020/05/11/is-building-neural-nlg-faster/>.
- Anna Rogers. 2019. How the transformers broke nlp leaderboards. Posted on the *Hacking Semantics* blog: <https://hackingsemantics.xyz/2019/leaderboards/>.
- David Schlangen. 2019. [Language tasks and language games: On methodology in current natural language processing research](#). *CoRR*, abs/1908.10747.
- David Schlangen. 2020. [Targeting the benchmark: On methodology in current natural language processing research](#). *CoRR*, abs/2007.04792.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and policy considerations for deep learning in NLP](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. [BLEU is not suitable for the evaluation of text simplification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.
- Pierre-Luc Vaudry and Guy Lapalme. 2013. [Adapting SimpleNLG for bilingual English-French realisation](#). In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, Sofia, Bulgaria. Association for Computational Linguistics.

Huiyuan Xie, Tom Sherborne, Alexander Kuhnle, and Ann Copestake. 2019. [Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity](#). In *Proceedings of Rigorous Evaluation of AI Systems 2019, collocated with The seventh AAIL Conference on Human Computation and Crowdsourcing*.

NUBIA: NeUral Based Interchangeability Assessor for Text Generation

Hassan Kane *
MIT
WL Research

Muhammed Yusuf Kocyigit *
Boston University
WL Research

Ali Abdalla
WL Research

Pelkins Ajanoh
Harvard University
WL Research

Mohamed Coulibali
Laval University
WL Research

Abstract

We present NUBIA, a methodology to build automatic evaluation metrics for text generation using only machine learning models as core components. A typical NUBIA model is composed of three modules: a neural feature extractor, an aggregator and a calibrator. We demonstrate an implementation of NUBIA showing competitive performance with state-of-the-art metrics used to evaluate machine translation and state-of-the-art results for image captions quality evaluation. In addition to strong performance, NUBIA models have the advantage of being modular and improve in synergy with advances in text generation models.

1 Introduction

Evaluation metrics play a central role in the machine learning community. They direct research efforts and define the state-of-the-art models. Unlike machine learning tasks such as classification and regression, text generation (i.e. machine translation, summarization, image captioning) is a nuanced task where the gold standard for quality evaluation is human assessment. However, this method of evaluation is expensive and time consuming.

As a complement, automatic metrics were designed to approximate human judgment of quality. A consequence of this unique setup is that the metrics themselves have to be frequently upgraded to reflect the dynamic progress of the field. However this has not happened and, while the text generation models have dynamically evolved, the metrics most commonly to assess model outputs used have not.

The two most common metrics used for evaluating similarity between candidate and reference texts are BLEU (Bilingual Evaluation Under-

study) (Papineni et al., 2002) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). Both approaches rely on counting the matching n-grams in the candidate text to n-grams in the reference text. The former is precision focused while the latter is recall focused.

These metrics have posed serious limitations and have already been criticized by the academic community (Reiter, 2018; Callison-Burch et al., 2006; Sulem et al., 2018; Novikova et al., 2017). In this work, we propose a methodology to build text generation evaluation metrics using deep learning models as core components.

An implementation of this methodology is then presented and tested in the domains of machine translation and image captioning quality estimation. For assessing the metric in the machine translation domain, we use the WMT 2017, 2018 and 2019 dataset.

We conduct further experiments showing that, without any additional fine-tuning, the same model used to assess machine translation quality outperforms existing metrics specifically designed to assess image captioning quality.

Beyond the promise of this methodology in terms of its ability to lead to metrics with high correlation to human judgment, NUBIA metrics can be constructed with any base architecture, perform well with only thousands of examples as supervision signal and are expected to improve continuously with future NLP advances.

2 Related Work

2.1 BLEU, ROUGE and n-gram matching approaches

BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have been used as the main evaluation methods in a variety of NLP tasks for almost two decades. BLEU is shown to better correlate with

Equal contribution. Correspondence to Hassan Kane: <hassanmohamed@alum.mit.edu>

human judgment when the hypothesis texts are bad as we can see in figure 2(c) and correlate weakly when the hypothesis texts are better. CIDEr is an image captioning metric that computes cosine similarity between tf-idf weighted n-grams (Vedantam et al., 2015). METEOR (Banerjee and Lavie, 2005) uses the harmonic mean of unigram precision and recall in combination with synonym matching and stemming along with word matching. While n-gram matching approaches are fast and simple to understand, this paradigm is limited in its ability to capture higher order semantic meaning.

The shortcomings of these methods have been widely criticised and studied. Reiter (2018), in his structured review of BLEU, finds a low correlation between BLEU and human judgment. Callison-Burch et al. (2006) examine BLEU in the context of machine translation and find that BLEU neither correlates with human judgment on adequacy (whether the hypothesis sentence adequately captures the meaning of the reference sentence) nor on fluency (the quality of language in the hypothesis sentence). Sulem et al. (2018) examine BLEU – in the context of text simplification – on grammaticality, meaning preservation and simplicity. They report a very low, and, in some cases, negative correlation with human judgment.

2.2 Transformers, BERT and GPT

Language modeling has become an important NLP technique, thanks to its ability to be applied to various NLP tasks as explained in (Radford et al., 2019). There are two leading architectures for language modeling: Recurrent Neural Networks (RNNs) (Mikolov et al., 2010) and Transformers (Vaswani et al., 2017). RNNs handle the input tokens, words or characters, one by one through time and learn the relationship between them, whereas transformers receive a segment of tokens and learn the dependencies between them using an attention mechanism.

The recent success of transformers as multitask learners (Radford et al., 2019) motivated us to adapt them for the task of neural language evaluation. This is crucial because what stood as an obstacle before neural language models was the power to generalize well to different datasets and tasks. Now with models like GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) trained on huge amounts of data, we can start trusting their ability to generalize across domains. As of now, machine

summarization, translation and image captioning all use different metrics to compare reference sentences with candidate sentences. Transformers-based models offer the promise to unify quality evaluation across tasks.

2.3 Model-based metrics

While BLEU and ROUGE are defined in a discrete space of word tokens, other evaluation metrics are powered by neural networks and word vectors. BERTscore (Zhang* et al., 2020) computes word embeddings and cosine similarity to create a score array and uses greedy matching to maximize the similarity score between words in the candidate and reference sentences. Sentence Mover’s Similarity (Clark et al., 2019) uses a Wasserstein metric defined on sentence embeddings generated from averaging the word embeddings in a sentence. YiSi (Lo, 2019) also defines a distance metric among reference and hypothesis sentences based on multilingual BERT embeddings and word frequency weightings. SPICE (Anderson et al., 2016) is an image captioning metric that creates a parse tree from the reference caption, candidate caption to create a scene graph and compute a score based on the overlapping relationships.

These methods report stronger correlations with human judgment and better results when compared with BLEU and ROUGE. While they are using word embeddings (Mikolov et al., 2013) to convert their sentences in a continuous space, they use hand-crafted mathematical functions to evaluate similarity in that space. In NUBIA, rather than defining a mathematical formula, we train a neural network to learn it using human judgement on thousands of sentence pairs as supervision signal.

BLEND (Ma et al., 2017) uses an SVM to combine different existing evaluation metrics. RUSE (Shimanaka et al., 2018) embeds both sentences separately and pools them to a given size. After, the method uses a pre-trained MLP to predict on different tasks. This quality estimator metric is then proposed to be used in language evaluation.

BLEURT (Sellam et al., 2020) introduces a BERT model in combination with a novel pre-training scheme that uses millions of synthetic examples to help the model generalize and then fine-tune it on human judgement.

Our proposed methodology is also a learned metric. Instead of synthesizing millions of examples, we use different pre-trained transformers as feature

extractors on reference, hypothesis sentence pairs and then learn a mapping between those features and a final quality score.

2.4 GLUE Benchmark

The GLUE Benchmark is a collection of tools for evaluating and analyzing the performance of NLP models across a diverse range of tasks (Wang et al., 2018). The recent introduction of this benchmark has encouraged the NLP community to move away from specialized models doing well on a single task to models performing well across diverse tasks. NLP models such as transformers are usually pre-trained on a large corpus in an unsupervised manner and then fine-tuned on a dataset used for the specific task of the benchmark. Architectures doing well on this benchmark can be used as components of future NUBIA models

3 NUBIA model

Our method has three modules: a neural feature extractor, an aggregator and a calibrator. The feature extractor tested in this paper consists of different transformer architectures fine-tuned on relevant tasks of language evaluation such as semantic similarity, logical inference and sentence likelihood. While we use these features and architectures as the main building blocks of NUBIA, the specific models can change as long as they maintain the necessary performance in terms of correlation with human judgment on the fine-tuning tasks.

The aggregator uses the features extracted by the transformers as well as non-neural features such as reference and candidate sentence length and is trained to predict the quality of the hypothesis sentence given the reference sentence. Similar to the WMT challenge, we use past years' data to train this aggregator and test it on the test subset.

The calibrator is the final module that caps all predictions to be between 0 and 1.

3.1 Neural Feature Extraction

In this section, we will describe how we broke down the problem of assessing the quality of a sentence into numerical features, the thought process behind the features used and provide details on the models used for one possible implementation of a NUBIA architecture.

3.1.1 Semantic similarity

The first feature extracted between candidate and reference sentence is semantic similarity. In our

proposed implementation, we use a RoBERTa large pre-trained model (Liu et al., 2019), which we fine-tune to predict sentence similarity (0-5 scale) on the STS-B benchmark dataset (8,628 sentence pairs).

The rationale for this feature is that a good candidate sentence should have high semantic similarity with the reference sentence.

3.1.2 Logical Entailment

The second set of features looks at the logical relationship between the reference and hypothesis sentence. The quality of the generated text depends not only on the grammar and semantics but also the core meaning and argument of the candidate sentence. A good model will output sentences that convey the same message.

To extract these features, we use a RoBERTa large pre-trained model (Liu et al., 2019) which is then fine-tuned on the MNLI challenge from the GLUE benchmark.

The MNLI model is trained to take as input sentence pairs and output 0 if the sentences are in contradiction with each other, 1 if the logical relationship is undetermined/neutral (i.e. sentences do not discuss the same topic) and 2 if the sentence are in logical agreement with each other.

We take the likelihood scores over the 3 possible classes as features.

3.1.3 Sentence Intelligibility

The third set of neural features aims to capture the linguistic acceptability of the candidate sentence.

The rationale of this feature is that we want to make sure that candidate sentences are legible and grammatically correct.

It is a common failure mode for machine translation models to generate sentences which are close in meaning to the reference sentence but introduce uncommon syntax and grammatical errors. We currently model this by using the perplexity score of a state-of-the-art Neural Language Model: GPT-2 (Radford et al., 2018)

More precisely, given a sentence A and a sentence B, the 2 features we compute are the perplexity scores for sentence A and sentence B. Optionally, in one of the NUBIA version, we also introduce the number of words in the candidate and reference sentences. We have experimentally found that adding these features in conjunction with the perplexity scores improves correlation with human judgment.

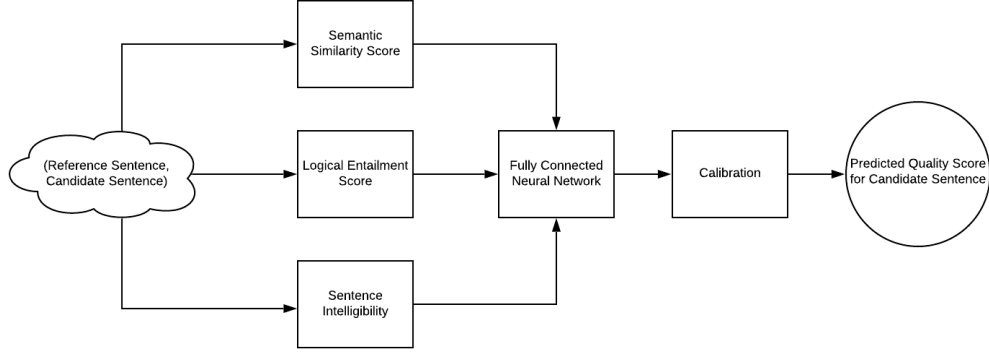


Figure 1: Outline of a NUBIA model with the three steps of Neural Feature Extraction, Aggregation and Calibration.

3.2 Aggregator

In the section above, we defined the dimensions used to assess the quality of a candidate sentence and then showed how to turn these dimensions into numerical scores using transformer models. The aggregator module is trained to approximate a function mapping input neural features to a quality score reflecting how interchangeable the candidate sentence and the reference sentences are.

The inspiration behind this model is that when human evaluators assess the quality of a candidate sentence against a reference sentence, they simultaneously pay attention to several aspects of the candidate sentence such as its semantic similarity with the reference sentence and whether it makes grammatical sense.

Since the relationship between these features and human judgement of quality is unknown, the goal of the aggregator is to approximate it using data obtained from rigorously conducted human evaluations.

The aggregator is a regression model trained to predict human evaluation on pairs of candidate and reference sentences. In this work, we explored linear regression and feed-forward, fully connected neural network architectures.

The neural network aggregator is a fully-connected, feedforward neural network architectures with either 6 (neural features only) or 8 (neural features and number of words in candidate and reference sentences) input layers corresponding to the features extracted, 10 hidden layers and a 1 dimension output layer corresponding to the human score prediction. The activation function for the model is the hyperbolic tangent and the optimizer

is ADAM (Kingma and Ba, 2015). NUBIA models using 6 input features have the "NUBIA-6DIM" prefix while the NUBIA models using 8 input features have the "NUBIA-8DIM" prefix. Models using a neural network as an aggregator have the "-NN" suffix while those using linear regression have the "-LREG" suffix.

3.3 Calibration

In practice, the output of the regressors are already highly correlated with human judgement; however, they lack two important properties. The first one is that the regressed score comparing a reference sentence with itself is not always equal to 1. To remedy to this, we normalize the scores given to a candidate sentence by the score given by the regressor of the candidate sentence with itself. The second missing property is that the raw regression scores are not strictly bounded to be between 0 and 1. To ensure they are, we cap the output of the regressors to have a value between 0 and 1.

4 Experiments

To assess our proposed implementation, we used both direct assessment and segment-level relative ranking from different WMT metrics shared tasks (Bojar et al., 2017; Graham et al., 2015) as well as tasks from image captioning. We did not conduct experiments in the domain of machine summarization because there are no labeled datasets containing pairs of summaries and their corresponding human evaluations of the summary quality.

In the WMT Direct Assessment task, candidate and reference translations are given for several language pairs and for each candidate translation, 15 human evaluators assign a quality score between

0 and 100. The final human score is taken as the average of the 15 human assessments. The performance of metrics is assessed using Pearson correlation with human judgement. For this task, we used the 2017 dataset because, unlike the WMT 2018 and WMT 2019 dataset, each sentence has been scored by at least 15 human evaluators (Ma et al., 2018).

For relative ranking, WMT 2018 and WMT 2019 still use direct human assessments but since there is not at least 15 annotators per sentence pairs, the direct assessment correlation task is converted into relative ranking task. More specifically, for a given reference sentences, up to 5 machine translation systems generate candidate translations. These candidate sentences are rated by human annotators on a discrete 0-25-50-75-100 points scale. After averaging the human annotations, if the gap between two candidate translation is higher than 25 points, one translation is considered to be better than the other. When the gap between two candidate sentences is lower than 25 points, the sentence pairs are not included in the segment-level evaluation Ma et al. (2018). In that setting, metrics are scored on their ability to preserve the human ranking using the Kendall’s Tau correlation coefficient.

4.1 Model training and testing

4.1.1 Machine Translation

For the machine translation experiments, we use the WMT 2015, 2016, 2017, 2018 and 2019 datasets in different settings. In these datasets, we only picked translations where the target language is English. This was done because the language models we used and their underlying word embeddings are trained on English sentences. All datasets are used for testing in future years.

For the WMT 2017 dataset (3,920 sentence pairs), we use an aggregator trained on human judgement from WMT 2015 and 2016 (5,360 sentence pairs). For the WMT 2018 (207,576 sentence pairs) and WMT 2019 (281,009 sentence pairs), we used an aggregator trained on WMT 2015 through 2017 (9,280 sentence pairs). In practice we found no improvement by adding sentences from WMT 2018 to train the aggregator which is why we stick with WMT 2015 through 2017 to test on both WMT 2018 and WMT 2019.

Feature extraction was conducted using one P100 GPU instance and took 3 hours for WMT 2017 and four days for WMT 2018 and 2019.

For the WMT 2017 task, the performance metric is Pearson correlation with human judgement. For the WMT 2018 and WMT 2019 challenges which are focused on relative ranking, metrics are compared with a Kendall’s Tau formulation on how well their scores correlate with human rankings of machine translation.

4.1.2 Image Captioning

For image captioning, we followed SPICE and used the Flickr 8K dataset. This dataset consists of 8,092 images annotated with 5 gold standard captions generated by humans. The dataset also has a human-evaluated part where for each image, a candidate caption is selected from the entire dataset and scored by three expert judges between 1 (“the selected caption is unrelated to the image”) and 4 (“the selected caption describes the image without any error.”). This part has 5,822 human-evaluated image caption pairs where each image also has 5 reference gold standard captions.

NUBIA is compared with Kendall’s Tau on how well it correlates with the average of the three judges’ scores as labels. Neural Feature extraction was conducted using one P100 GPU instance and took 12 hours. The aggregators for the NUBIA models used in the image captioning experiments are not specifically fine-tuned for the task and consist of the Neural Feature Extractors described above along with an aggregator trained on the WMT 2015, WMT 2016 and WMT 2017 dataset (9,280 sentence pairs).

5 Results

In Table 2, we report our results on the test set. We compare our methods with methods developed for the WMT 2017 challenge and recent models like BERTScore and BLEURT which are currently the best performing methods. Although many methods have been proposed throughout the years in the WMT metrics challenge, the current methods used to this day to assess performance of machine translation models are still BLEU and ROUGE score. For ROUGE, we use ROUGE-L scores because it is the formulation of ROUGE correlated the most with human judgements on WMT 2017.

In Table-3, we report the results for the relative ranking test of WMT 2018. Here we see that NUBIA is only outperformed by BLEUR. In Table-4, we have the results for the WMT 2019 challenge. Here we observe that NUBIA performs comparably with other methods.

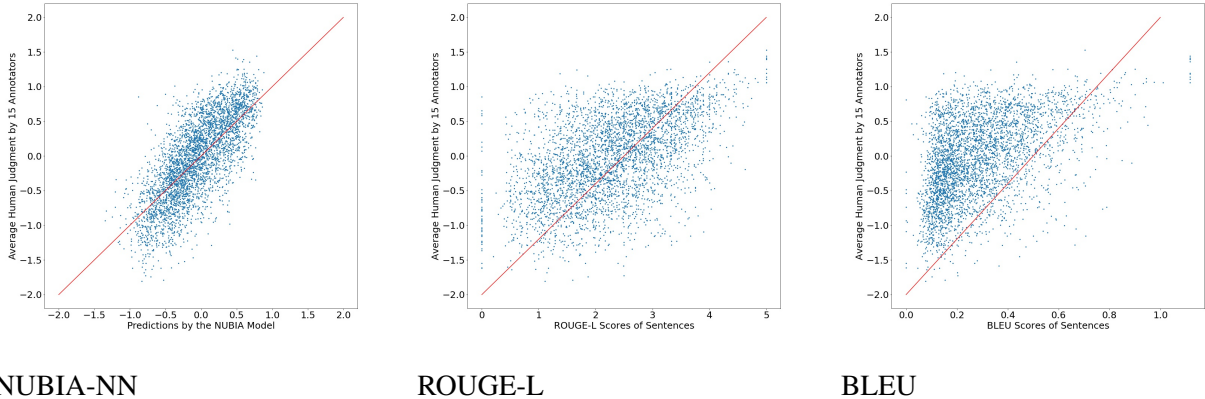


Figure 2: Score and label graphs of NUBIA, ROUGE-L and BLEU for the entire WMT-2017 segment level sets.

Model	τ
BLEU-1*	0.32
BLEU-4	0.33
ROUGE-L*	0.32
BERTScore	0.394
METEOR*	0.42
BLEURT	0.434
CIDEr*	0.44
SPICE*	0.45
NUBIA-6DIM-NN	0.47
NUBIA-8DIM-NN	0.495

Table 1: Kendall’s Tau Correlation with human judgment on Flickr 8K dataset. The scores marked with * are taken directly from the original SPICE paper. The BLEU-4 score in the original paper was 0.14 but the experiment was repeated with a smoothed function and the new result is reported.

We report the results of the image captioning experiments in Table-1. Here we observe that NUBIA outperforms all existing methods and achieves state-of-the-art correlation with human judgment of caption quality.

The strong performance maintained across varied tasks is a strong indicator of the robustness of this methodology and shows its promise to generalize well beyond the training set.

5.1 Ablation Study

To judge the importance of the features we have picked, we ran an ablation study where we trained a NUBIA model with only a subset of the features and report correlation results on the WMT17 dataset. The most crucial feature is the RoBERTa semantic similarity score. As suspected, other elements beyond semantic similarity also seem to be factored into prediction of translation quality as evidenced by the performance boost obtained after computing the GPT-2 features and MNLI features.

5.2 Error Analysis

Figure 2 sheds more light on the behavior of BLEU and ROUGE, two of the most common evaluation metrics and NUBIA-NN. This analysis unveils important properties of these metrics and helps better understand their strengths and weaknesses.

If we start with (c) we can see that BLEU correlates better with human judgment in the bottom left (bad hypothesis area). Essentially, if a human is likely to give a bad score to a sentence, BLEU is unlikely to overscore. But if a person is going to give a high score, BLEU is equally likely to give any score, maybe even more likely to penalize the sentence. This effectively inhibits the desired behaviour in language generation.

While the behaviour of ROUGE is much more balanced, it is still prone to underscoring and overscoring.

When we look at NUBIA-NN, we see a general trend followed along the data, as expected given the

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	r	r	r	r	r	r	r	r
BLEU	0.432	0.425	0.577	0.415	0.479	0.548	0.515	0.484
ROUGE-L	0.482	0.492	0.623	0.465	0.480	0.593	0.569	0.529
BLEND	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.632
MEANT2.0	0.578	0.565	0.687	0.586	0.607	0.596	0.639	0.608
RUSE	0.614	0.637	0.756	0.705	0.680	0.704	0.677	0.681
NUBIA-6DIM-LReg	0.739	0.733	0.815	0.788	0.734	0.766	0.763	0.763
NUBIA-8DIM-LReg	0.739	0.732	0.829	0.783	0.731	0.784	0.768	0.767
BERTscore	0.714	.740	0.835	0.774	0.773	0.776	0.767	0.768
NUBIA-6DIM-NN	0.745	0.730	0.847	0.779	0.737	0.800	0.751	0.770
NUBIA-8DIM-NN	0.754	.738	0.854	0.786	0.755	0.804	0.750	0.777
BLEURT	0.773	0.792	0.878	0.835	0.811	0.824	0.814	0.818

Table 2: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	τ	τ	τ	τ	τ	τ	τ	τ
BLEU	0.268	0.458	0.311	0.206	0.259	0.178	0.210	0.270
ROUGE-L	0.28	0.473	0.324	0.208	0.275	0.193	0.211	0.281
YiSi1 SRL 18	0.317	0.483	0.345	0.237	0.306	0.233	0.209	0.304
RUSE	0.3478	0.498	0.368	0.273	0.311	0.259	0.218	0.325
YiSi1 SRL 19	0.396	0.543	0.39	0.303	0.351	0.297	0.253	0.362
Yisi1	0.391	0.544	0.397	0.299	0.352	0.301	0.254	0.363
BERTScore	0.408	0.550	0.395	0.293	0.346	0.296	0.260	0.364
NUBIA-6DIM-NN	0.396	0.550	0.410	0.326	0.357	0.295	0.262	0.371
NUBIA-8DIM-NN	0.402	0.553	0.410	0.330	0.357	0.288	0.268	0.373
BLEURT	0.423	0.567	0.414	0.325	0.360	0.315	0.260	0.381

Table 3: Kendall’s Tau correlation with segment-level human judgments on WMT18 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

high correlation score. The only interesting action is the overscoring of low human score sentences. The nature of the error can be analyzed to further improve NUBIA.

6 Conclusion

In this work, we introduced NUBIA: a methodology to build automatic evaluation metrics for text generation using machine learning models as core components. An implementation of this methodology achieves strong results on machine translation and state-of-the-art results on image captioning strongly building on the successes of recent NLP architectures such as RoBERTa and GPT-2. These strong results are achieved using a small amount of supervised training data. This methodology offers the possibility of building evaluation metrics im-

proving in synergy with the progress of generative models and unifying evaluation of image captioning, machine translation and potentially other text generation tasks.

7 Discussion and future work

Learned text generation evaluation metrics have enormous promise to change how text generation models are assessed. Future work can further probe which other text generation tasks NUBIA models are strong candidates to assess.

NUBIA can be improved along four axes. The first axis of improvement is through the efforts of the wider NLP community at creating models achieving strong results on the NLU benchmarks like GLUE. The second axis is through the addition of better features capturing aspects of human

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	τ	τ	τ	τ	τ	τ	τ	τ
BLEU	0.173	0.264	0.207	0.389	0.280	0.166	0.349	0.261
ROUGE-L	0.169	0.268	0.198	0.394	0.294	0.171	0.348	0.263
ESIM	0.167	0.337	0.303	0.435	0.359	0.201	0.396	0.314
NUBIA-6DIM-NN	0.248	0.356	0.274	0.419	0.385	0.227	0.410	0.331
YISI	0.199	0.346	0.306	0.442	0.380	0.222	0.431	0.332
NUBIA-8DIM-NN	0.251	0.358	0.258	0.429	.385	0.229	0.413	0.332
BERTscore	0.230	0.345	0.320	0.432	0.381	0.223	0.444	0.339
BLEURT	0.169	0.363	0.319	0.446	0.406	0.223	0.424	0.336

Table 4: Kendall’s Tau correlation with segment-level human judgments on WMT19 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	r	r	r	r	r	r	r	r
NUBIA-NN,SI	0.412	0.451	0.624	0.571	0.447	0.437	0.410	0.478
NUBIA-NN,LI	0.620	0.539	0.693	0.647	0.603	0.692	0.571	0.623
NUBIA-NN,LI+SI	0.643	0.621	0.775	0.722	0.646	0.681	0.624	0.673
NUBIA-NN,SS	0.678	0.686	0.790	0.740	0.694	0.766	0.708	0.723
NUBIA-NN,SS+LI	0.696	0.699	0.804	0.758	0.708	0.784	0.723	0.738
NUBIA-NN,SS+SI	0.727	0.729	0.842	0.785	0.726	0.790	0.755	0.764
NUBIA-NN,SS+LI+SI	0.754	0.738	0.854	0.786	0.755	0.804	0.750	0.777

Table 5: Ablation study results for NUBIA-NN on WMT 2017 Direct Assessment task. SS=Semantic Similarity, LI=Linguistic Inference, SI=Sentence Intelligibility.

quality assessment. Two candidate features are the linguistic acceptability which can be obtained by using models trained on the CoLA challenge and a coherence score for long text generations. The third axis is through better aggregator design. Finally, the fourth axis is reducing the computational cost of NUBIA models. The transformer architectures used as backbone for feature extraction are currently independent of each other. Using lighter models or fine-tuning using shared layers could lead to less compute-intensive models.

Learning how to specify NUBIA architectures and standardizing nomenclature will be crucial to ensure adoption, reproducibility and fair comparison of models scored using such automatic metrics. An exhaustive solution can be to describe the individual feature extractor. This description should not only include architectures but also training data and fine-tuning data (Mitchell et al., 2019; Gebru et al., 2018; Bender and Friedman, 2018). Similarly, aggregators should also be described through their architectures along with the training corpus

Evaluation and scorecards for neural metrics

going beyond correlation with human judgement (Boag et al., 2016) will help shed lights on their inner workings and failure modes. Such setups more precisely measure the effect that systematic sentence transformations (e.g. active to passive voice) have on the automatic metric scores.

Closely related to evaluation and data reporting, biased training data leading to underscoring or over scoring of valid translations should also be investigated.

Another area of current limitation is the language. Existing NUBIA models only work for English sentence pairs though the procedure to generate and assess such metrics in other languages is likely to be similar.

Understanding how such models can be adversarially attacked is also an open research question.

Finally, future work can also investigate convergence behavior and output of training setups where NUBIA is used as a loss function of text generation models.

References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- William Boag, Renan Campos, Kate Saenko, and Anna Rumshisky. 2016. **MUTT: Metric unit TesTing for language generation tasks**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1935–1943, Berlin, Germany. Association for Computational Linguistics.
- Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. **Results of the WMT17 metrics shared task**. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. **Re-evaluating the role of Bleu in machine translation research**. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. **Sentence mover’s similarity: Automatic evaluation for multi-sentence texts**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
- Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. **Accurate evaluation of segment-level machine translation metrics**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Chi-kiu Lo. 2019. **YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.
- Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. **Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance**. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.
- Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. **Blend: a novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task**. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603, Copenhagen, Denmark. Association for Computational Linguistics.
- Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.
- Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. **Why we need new evaluation metrics for NLG**. *CoRR*, abs/1707.06875.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Ehud Reiter. 2018. [A structured review of the validity of BLEU](#). *Computational Linguistics*, 44(3):393–401.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.
- Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. [RUSE: Regressor using sentence embeddings for automatic machine translation evaluation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.
- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. [BLEU is not suitable for the evaluation of text simplification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- 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*.

On the interaction of automatic evaluation and task framing in headline style transfer

Lorenzo De Mattei^{*◊†}, Michele Cafagna[‡], Huiyuan Lai[†],
Felice Dell’Orletta[◊], Malvina Nissim[†], Albert Gatt[‡]

^{*} Department of Computer Science, University of Pisa / Italy

[◊] ItaliaNLP Lab, Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa / Italy

[†] CLCG, University of Groningen / The Netherlands

[‡] LLT, University of Malta / Malta

lorenzo.demattei@di.unipi.it

{michele.cafagna, albert.gatt}@um.edu.mt

{h.lai, m.nissim}@rug.nl

felice.dellorletta@ilc.cnr.it

Abstract

An ongoing debate in the NLG community concerns the best way to evaluate systems, with human evaluation often being considered the most reliable method, compared to corpus-based metrics. However, tasks involving subtle textual differences, such as style transfer, tend to be hard for humans to perform. In this paper, we propose an evaluation method for this task based on purposely-trained classifiers, showing that it better reflects system differences than traditional metrics such as BLEU and ROUGE.

1 Introduction and Background

The evaluation of Natural Language Generation (NLG) systems is intrinsically complex. This is in part due to the virtually open-ended range of possible ways of expressing content, making it difficult to determine a ‘gold standard’ or ‘ground truth’. As a result, there has been growing scepticism in the field surrounding the validity of corpus-based metrics, primarily because of their weak or highly variable correlations with human judgments (Reiter and Sripada, 2002; Reiter and Belz, 2009; Reiter, 2018; Celikyilmaz et al., 2020). Human evaluation is generally viewed as the most desirable method to assess generated text (Novikova et al., 2018; van der Lee et al., 2019). In their recent comprehensive survey on the evaluation of NLG systems, Celikyilmaz et al. (2020) stress that it is important that any used untrained automatic measure (such as BLEU, ROUGE, METEOR, etc) correlates well with human judgements.

At the same time, human evaluation also presents its challenges and there have been calls

for the development of new, more reliable metrics (Novikova et al., 2017). Beyond the costs associated with using humans in the loop during development, it also appears that certain linguistic judgment tasks are hard for humans to perform reliably. For instance, human judges show relatively low agreement in the presence of syntactic variation (Cahill and Forst, 2009). By the same token, Dethlefs et al. (2014) observe at best moderate correlations between human raters on stylistic dimensions such as politeness, colloquialism and naturalness.

Closer to the concerns of the present work, it has recently been shown that humans find it difficult to identify subtle stylistic differences between texts. De Mattei et al. (2020b) presented three independent judges with headlines from two Italian newspapers with distinct ideological leanings and in-house editorial styles. When asked to classify the headlines according to which newspaper they thought they came from, all three annotators performed the task with low accuracy (ranging from 57% to 62%). Furthermore, agreement was very low (Krippendorff’s $\alpha = 0.16$). Agreement was similarly low on classifying automatically generated headlines ($\alpha = 0.13$ or 0.14 for two different generation settings). These results suggest that human evaluation is not viable, or at least not sufficient, for this task.

In this work we focus on the same style-transfer task using headlines from newspapers in Italian, but address the question of whether a series of classifiers that monitor both style strength as well as content preservation, the core aspects of style transfer (Fu et al., 2018; Mir et al., 2019; Luo et al.,

2019), can shed light on differences between models.

We also add some untrained automatic metrics for evaluation. As observed above, the fact that humans cannot perform this task reliably makes it impossible to choose such metrics based on good correlations with human judgement (Celikyilmaz et al., 2020). Therefore, relying on previous work, we compare the insights gained from our classifiers with those obtained from BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), since they are commonly used metrics to assess performance for content preservation and summarisation. Other common metrics such as METEOR (Banerjee and Lavie, 2005) and BLEURT (Sellam et al., 2020), which in principle would be desirable to use, are not applicable to our use case as they require resources not available for Italian.

More specifically, we train a classifier which, given a headline coming from one of two newspapers with distinct ideological leanings and in-house styles, can identify the provenance of the headline with high accuracy. We use this (the ‘main’ classifier) to evaluate the success of a model in regenerating a headline from one newspaper, in the style of the other. We add two further consistency checks, both of which aim at content assessment, and are carried out using additional classifiers trained for the purpose: (a) a model’s output headline should still be compatible in content with the original headline; (b) the output headline should also be compatible in content with the article to which it pertains. A headline is deemed to be (re)generated successfully in a different style if both (a) and (b) are satisfied, and the main classifier’s decision as to its provenance should be reversed, relative to its decision on the original headline.

A core element in our setup is testing our evaluation classifiers/strategies in different scenarios that arise from different ways of framing the style transfer task, and different degrees of data availability. Indeed, we frame the task either as a translation problem, where a headline is rewritten in the target style or as a summarisation problem, where the target headline is generated starting from the source article, using a summarisation model trained on target style. The two settings differ in their needs in terms of training data as well as in their ability to perform the two core aspects of style transfer (style strength and content preservation).

We observe how evaluation is affected by the

different settings, and how this should be taken into account when deciding what the best model is.

Data and code used for this paper are available at <https://github.com/michelecafagna26/CHANGE-IT>. The data and task settings also lend themselves well as material for a shared task, and they have indeed been used, with the summarisation system described here as baseline, in the context of the EVALITA 2020 campaign for Italian NLP (De Mattei et al., 2020a).

2 Task and Data

Our style transfer task can be seen as a “headline translation” problem. Given a collection of headlines from two newspapers at opposite ends of the political spectrum, the task is to change all rightwing headlines to headlines with a leftwing style, and all leftwing headlines to headlines with a rightwing style, while preserving content. We focus on Italian in this contribution, but the methodology we propose is obviously applicable to any language for which data is available.

Collection We used a corpus of around 152K article-headline pairs from two wide circulation Italian newspapers at opposite ends of the political spectrum namely *la Repubblica* (left-wing) and *Il Giornale* (right-wing) provided by De Mattei et al. (2020b). The data is balanced across the two sources. Though we are concerned with headlines, full articles are used in two ways: (a) *alignment*; and (b) the consistency check classifiers (see Section 4 for details). For the former, we leverage the alignment procedure proposed by Cafagna et al. (2019) and we split our dataset into strongly aligned, weakly aligned and non-aligned news. The purpose of alignment is to control for potential topic biases in the two newspapers so as to better disentangle newspaper-specific style. Additionally, this information is useful in the creation of our datasets, specifically as it addresses the need for parallel data for our evaluation classifiers and the translation-based model (see below).

Alignment We compute the tf-idf vectors of all the articles of both newspapers and create subsets of relevant news filtering by date, i.e. considering only news which were published approximately within the same, short time interval for the two sources. On the tf-idf vectors we then compute cosine similarities for all news in the resulting subset, rank them, and retain only the alignments that are

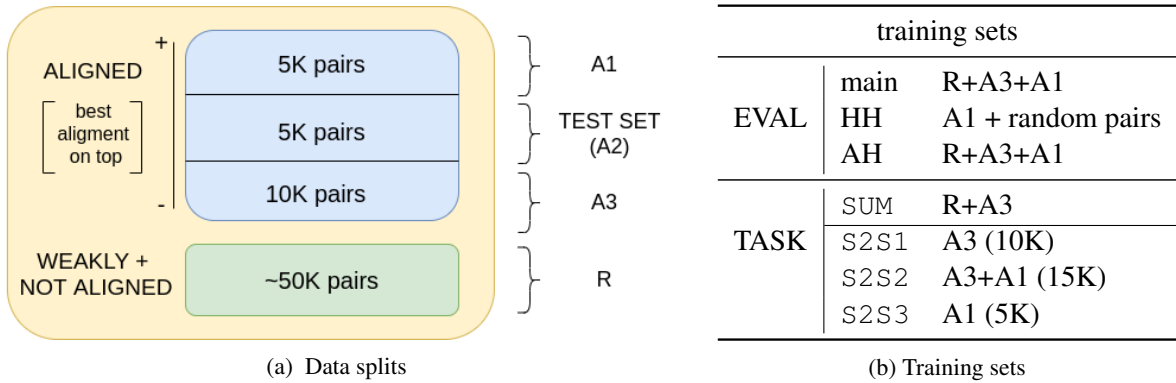


Figure 1: Data splits and their use in the different training sets

above a certain threshold. The threshold is chosen taking into consideration a trade-off between number of documents and quality of alignment. We choose two different thresholds: one is stricter (> 0.5) and we use it to select the best alignments; the other one is looser (> 0.185 , and ≤ 0.5).

Data splitting We split the dataset into *strongly aligned news*, which are selected using the stricter threshold ($\sim 20K$ aligned pairs), and *weakly aligned and non-aligned news* ($\sim 100K$ article-headline pairs equally distributed among the two newspapers). The aligned data is further split as shown in Figure 1a. SA is left aside and used as test set for the final style transfer task. The remaining three sets are used for training the evaluation classifiers and the models for the target task in various combinations. These are described in Figure 1b and in connection with the systems’ descriptions.¹

3 Systems

Our focus is on the interaction of different evaluation settings and approaches to the task. Accordingly, we develop two different frameworks with different takes on the same problem: (a) as a true translation task, where given a headline in one style, the model learns to generate a new headline in the target style; (b) as a summarisation task, where headlines are viewed as an extreme case of summarisation and generated from the article. We exploit article-headline generators trained on opposite sources to do the transfer. This approach does not in principle require parallel data for training.

For the translation approach (S2S), we train a supervised BiLSTM sequence-to-sequence model with attention from OpenNMT (Klein et al., 2017)

¹Note that all sets also always contain the headlines’ respective full articles, though these are not necessarily used.

to map the headline from left-wing to right-wing, and viceversa. Since the model needs parallel data, we exploit the aligned headlines for training. We experiment with three differently composed training sets, varying not only in size, but also in the strength of the alignment, as shown in Figure 1b.

For the summarisation approach (SUM), we use two pointer-generator networks (See et al., 2017), which include a *pointing mechanism* able to copy words from the source as well as pick them from a fixed vocabulary, thereby allowing better handling of out-of-vocabulary words. ability to reproduce novel words. One model is trained on the *la Repubblica* portion of the training set, the other on *Il Giornale*. In a style transfer setting we use these models as follows: Given a headline from *Il Giornale*, for example, the model trained on *la Repubblica* can be run over the corresponding article from *Il Giornale* to generate a headline in the style of *la Repubblica*, and vice versa. To train the models we use subset R, but we also include the lower end of the aligned pairs (A3), see Figure 1b.

4 Evaluation

Our fully automatic strategy is based on a series of classifiers to assess style strength and content preservation. For style, we train a single classifier (*main*). For content, we train two classifiers that perform two ‘consistency checks’: one ensures that the two headlines (original and transformed) are still compatible (*HH classifier*); the other ensures that the headline is still compatible with the original article (*AH classifier*). See also Figure 1a.

In what follows we describe these classifiers in more detail. When discussing results, we will show how the contribution of each classifier is crucial towards a comprehensive evaluation.

Main classifier The main classifier uses a pre-trained BERT encoder with a linear classifier on top fine-tuned with a batch size of 256 and sequences truncated at 32 tokens for 6 epochs with learning rate 1e-05. Given a headline, this classifier can distinguish the two sources with an f-score of approximately 80% (see Table 1). Since style transfer is deemed successful if the original style is lost in favour of the target style, we use this classifier to assess how many times a style transfer system manages to reverse the main classifier’s decisions.

HH classifier This classifier checks compatibility between the original and the generated headline. We use the same architecture as for the main classifier with a slightly different configuration: max. sequence length of 64 tokens, batch size of 128 for 2 epochs (early-stopped), with learning rate 1e-05. Being trained on strictly aligned data as positive instances (A1), with a corresponding amount of random pairs as negative instances, it should learn whether two headlines describe the same content or not. Performance on gold data is .96 (Table 1).

AH classifier This classifier performs yet another content-related check. It takes a headline and its corresponding article, and tells whether the headline is appropriate for the article. The classifier is trained on article-headline pairs from both the strongly aligned and the weakly and non-aligned instances (R+A3+A1, Figure 1b). At test time, the generated headline is checked for compatibility against the source article. We use the same base model as for the main and HH classifiers with batch size of 8, same learning rate and 6 epochs. Performance on gold data is >.97 (Table 1).

		prec	rec	f-score
main	rep	0.77	0.83	0.80
	gio	0.84	0.78	0.81
HH	match	0.98	0.95	0.96
	no match	0.95	0.98	0.96
AH	match	0.96	0.99	0.98
	no match	0.99	0.96	0.97

Table 1: Performance of the classifiers on gold data.

Overall compliancy We calculate a compliancy score which assesses the proportion of times the following three outcomes are successful (i) the *HH classifier* predicts ‘match’; (ii) the *AH classifier* predicts ‘match’; (iii) the *main classifier*’s decision is *reversed*. As upperbound, we find the compati-

bility score for gold at 74.3% for transfer from *La Repubblica* to *Il Giornale* (*rep2gio*), and 78.1% for the opposite direction (*gio2rep*).

5 Results and Discussion

Table 2 reports results of our evaluation methods both for the summarization system (SUM) and for the style transfer systems (S2S) in the different training set scenarios.

The top panel in Table 2 shows the results for systems where training data is weakly aligned or unaligned. The summarisation system SUM does better at content preservation (HH and AH) than S2S1. However, its scores on the *main* classifier are worse in both transfer directions, as well as on average. The average compliancy score is higher for S2S1. In summary, for data which is not strongly aligned, our methods suggest that style transfer is better when conceived as a translation task. BLEU is higher for SUM, but the overall extremely low scores across the board suggest that it might not be a very informative metric for this setup, although commonly used to assess content preservation in style transfer (Rao and Tetreault, 2018). Our HH and AH classifiers appear more indicative in this respect, and ROUGE scores seem to correlate a bit more with them, when compared to BLEU. It remains to be investigated whether BLEU, ROUGE, and our content-checking classifiers do in fact measure something similar or not.

With better-aligned data (bottom panel), the picture is more nuanced. Here, the main comparison is between two systems trained on strongly aligned data, one of which (S2S2) has additional, weakly aligned data. The overall compliancy score suggests that this improves style transfer (and this system is also the top performing one over all, also outperforming S2S1 and SUM). As for content preservation (AH and HH scores), S2S3 is marginally better on average for HH, but not for AH, where the two systems are tied.

Overall, the results of the classification-based evaluation also highlight a difference between a summarisation-based system (SUM), which tends to be better at content preservation, compared to a translation-based style transfer setup (especially S2S2) which transfers style better. Clearly, a corpus-based metric such as BLEU fails to capture these distinctions, but here does not appear informative even just for assessing content preservation.

		HH	AH	Main	Compl.	BLEU	ROUGE
without top aligned data							
SUM	rep2gio	.649	.876	.799	.449	.020	.145
	gio2rep	.639	.871	.435	.240	.026	.156
	avg	.644	.874	.616	.345	.023	.151
S2S1	rep2gio	.632	.842	.815	.436	.011	.136
	gio2rep	.444	.846	.864	.321	.012	.130
	avg	.538	.844	.840	.379	.012	.133
with top aligned data							
S2S2	rep2gio	.860	.845	.845	.549	.018	.159
	gio2rep	.612	.846	.847	.442	.016	.151
	avg	.736	.846	.849	.496	.017	.155
S2S3	rep2gio	.728	.844	.845	.520	.012	.139
	gio2rep	.760	.848	.649	.420	.013	.156
	avg	.744	.846	.747	.470	.013	.148

Table 2: Performance on test data.

One aspect that will require further investigation, since we do not have a clear explanation for it as of now, is the performance difference between the two translation directions. Indeed, transforming a *La Repubblica* headline into a *Il Giornale* headline appears more difficult than transforming headlines in the opposite directions, under most settings.

6 Conclusions

This paper addressed the issue of how to evaluate style transfer. We explicitly compared systems in terms of the extent to which they preserve content, and their success at transferring style. The latter is known to be hard for humans to evaluate (Dethlefs et al., 2014; De Mattei et al., 2020b). Our aim was primarily to see to what extent different evaluation strategies based on purposely trained classifiers could distinguish between models, insofar as they perform better at either of these tasks and in different training scenarios.

Our findings suggest that our proposed combination of classifiers focused on both content and style transfer can potentially help to distinguish models in terms of their strengths. Interestingly, a commonly used metric such as BLEU does not seem to be informative in our experiments, not even for the content preservation aspects.

To the extent that stylistic distinctions remain hard for humans to evaluate in setups such as the one used here, a classification-based approach with consistency checks for content preservation is a

promising way forward, especially to support development in a relatively cheap and effective way.

Future work will have to determine how the various metrics we have used relate to each other (especially our classifiers and BLEU/ROUGE), and whether human judgement can be successfully brought back, and in case in what form, at some stage of the evaluation process.

References

- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. ACL.
- Michele Cafagna, Lorenzo De Mattei, and Malvina Nissim. 2019. [Embeddings shifts as proxies for different word use in italian newspapers](#). In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy.
- Aoife Cahill and Martin Forst. 2009. [Human evaluation of a German surface realisation ranker](#). In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 112–120, Athens, Greece. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of text generation: A survey](#). *arXiv preprint arXiv 2006.14799*.

- Lorenzo De Mattei, Michele Cafagana, Felice Dell’Orletta, Malvina Nissim, and Albert Gatt. 2020a. **CHANGE-IT @ EVALITA 2020: Change Headlines, Adapt News, GEnerate**. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, and Malvina Nissim. 2020b. **Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment**. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France. European Language Resources Association.
- Nina Dethlefs, Heriberto Cuayáhuatl, 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*, pages 702–711, Gothenburg, Sweden. Association for Computational Linguistics.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. **Style transfer in text: Exploration and evaluation**. In *Proceedings of the Thirtieth Conference on Innovative Applications of Artificial Intelligence (IAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. **OpenNMT: Open-source toolkit for neural machine translation**. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Kraemer. 2019. **Best practices for the human evaluation of automatically generated text**. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. **A dual reinforcement learning framework for unsupervised text style transfer**. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. **Evaluating style transfer for text**. 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 495–504, Minneapolis, Minnesota. 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.
- 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 (EMNLP’17)*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Sudha Rao and Joel Tetreault. 2018. **Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer**. *arXiv preprint arXiv:1803.06535*.
- Ehud Reiter. 2018. **A Structured Review of the Validity of BLEU**. *Computational Linguistics*, 44(3):393–401.
- Ehud Reiter and Anja Belz. 2009. **An investigation into the validity of some metrics for automatically evaluating natural language generation systems**. *Computational Linguistics*, 35(4):529–558.
- Ehud Reiter and Somayajulu Sripada. 2002. **Should corpora texts be gold standards for NLG?** In *Proceedings of the International Natural Language Generation Conference*, pages 97–104, Harriman, New York, USA. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. **Get to the point: Summarization with pointer-generator networks**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. 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.

