

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 judge ¹ 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:

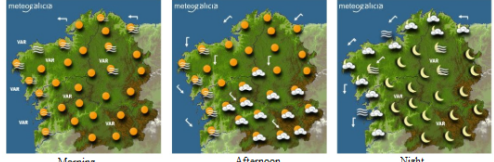
Situation 1



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.

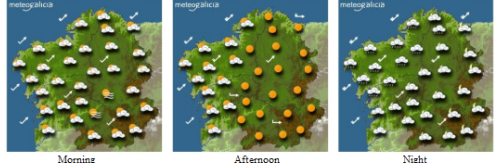
Situation 2



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. Civile. 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.