Evaluating the text quality, human likeness and tailoring component of PASS: A Dutch data-to-text system for soccer

Chris van der Lee
Tilburg University
c.vdrlee@tilburguniversity.edu

Bart Verduijn
Tilburg University
b.w.verduijn@tilburguniversity.edu

Emiel Krahmer
Tilburg University
e.j.krahmer@tilburguniversity.edu

Sander Wubben
Tilburg University
s.wubben@tilburguniversity.edu

Abstract

We present an evaluation of PASS, a data-to-text system that generates Dutch soccer reports from match statistics which are automatically tailored towards fans of one club or the other. The evaluation in this paper consists of two studies. An intrinsic human-based evaluation of the system's output is described in the first study. In this study it was found that compared to human-written texts, computer-generated texts were rated slightly lower on style-related text components (fluency and clarity) and slightly higher in terms of the correctness of given information. Furthermore, results from the first study showed that tailoring was accurately recognized in most cases, and that participants struggled with correctly identifying whether a text was written by a human or computer. The second study investigated if tailoring affects perceived text quality, for which no results were garnered. This lack of results might be due to negative preconceptions about computer-generated texts which were found in the first study.

1 Introduction

Evaluation of end-to-end Natural Language Generation (NLG) systems is important to assess whether the system has properly expressed certain properties (e.g. quality, speed), or whether the designed properties work as intended (Gkatzia and Mahamood, 2015). Traditional NLG evaluation approaches can typically be assigned to one of two categories: intrinsic or extrinsic (Belz and Reiter, 2006). Intrinsic approaches seek to evaluate properties of the system itself. This can be done using automatic measures such as BLEU, NIST, ROUGE, etc. or by asking human participants to rate the systems output with e.g. Likert or rating scales. Extrinsic approaches aim to assess the impact of the system, by measuring if the system can fulfill its purpose or what the user gains from the systems output.

While scholars have been positive about the effort the NLG community has put into their evaluations (Gatt and Belz, 2010, for instance), it is often the case that an extensive evaluation does not take much priority after a system is built. The usage of automatic measures is gaining traction due to its quickness and low costs, but they are still considered controversial by many (Reiter and Belz, 2009; Novikova et al., 2017, for instance). Furthermore, the intrinsic approaches with human ratings are often limited in scope, using a relatively small sample of participants, using relatively short questionnaires that only shine light on a small aspect of text quality, and/or comparing the computer-generated texts against non-representative human texts. Similarly, the amount of extrinsic evaluations that have been performed up until now is low. While they are considered the most useful type of evaluation by some (Reiter and Belz, 2009), they are not carried out as often as other types of evaluation.

In many cases, an extrinsic evaluation is also difficult because of the system's set-up. A system that is designed to produce texts that stay close to the facts and which purpose is to merely inform readers, should mainly be evaluated on the factual correctness of its output, which is covered by doing an intrinsic evaluation. A different kind of system was chosen as the object of evaluation in the current study: the Personalized Automated Soccer texts System (hereafter: PASS). This system generates summaries of soccer matches based on match statistics (van der Lee et al., 2017). A unique feature of PASS is that

Proceedings of the 27th International Conference on Computational Linguistics, pages 962–972
Santa Fe, New Mexico, USA, August 20-26, 2018.

the system produces texts tailored towards fans of one club or the other. Previously, scholars have emphasized that one of the potential strengths of data-to-text systems is their potential to produce multiple variants of texts based on the same data (Gatt and Krahmer, 2018). This tailoring has been suggested to improve attitudinal and behavioral outcomes in the context of human-written persuasive texts (Noar et al., 2007), but if these positive outcomes can also be found for computer-generated texts and/or texts with aims other than persuasion warrants further research.

PASS is relatively unique in the NLG landscape, because of its aim to not only inform but also to produce a text that is 'enjoyable' to read by its target audience, which it tries to accomplish by tailoring texts towards the wishes of an audience. These goals make a brief check for readability of its texts inadequate, and make a more extensive evaluation project necessary. The current study examined whether the PASS system succeeds in the aim to produce texts of significant linguistic quality and if the tailoring component plays a role in this aim, similar to how tailoring improves attitudinal and behavioral responses for human-written persuasive texts. To examine this, the goals of this paper were twofold. The first goal was to assess how the quality of PASS-generated texts fares against similar human texts, and the second goal was to assess whether the intended tailoring was clear and if it had effects on attitude towards the text. Besides these two main points, differences in preconceptions about human-written and computer-generated texts were also investigated, because these preconceptions could moderate the effectiveness of tailoring.

2 Background

2.1 Evaluation in NLG

Evaluation has become an increasingly important topic within the NLP domain as a whole, but also within the NLG domain more specifically. While the NLG domain has a strong evaluation tradition (Gatt and Belz, 2010), there is an ongoing discussion regarding the type of evaluation that should be used. One popular evaluation method is the use of metrics that can be computed automatically, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and NIST (Doddington, 2002). However, previous literature suggests that the use of such metrics should be done with caution. These metrics are attractive because they are quick, fast and repeatable (Reiter and Belz, 2009), and have in some cases been found to correlate with human judgments (Gkatzia and Mahamood, 2015). However, most if not all of the current automated metrics that are used in the NLG domain are based on overlap with a certain reference text. Therefore, it can be derived that such metrics are only useful if an aligned NLG output-reference text corpus can be constructed and if it can be assumed that qualitatively good output has a strong overlap with this standard. This will not be the case in many situations. Furthermore, these metrics also rely on the assumption that the reference text is a good representation of the 'best case scenario, which it often is not (Reiter and Sripada, 2002). Finally, it has been argued that such metrics do not provide a good assessment of many linguistic properties such as content selection, information structure, appropriateness, etcetera (Scott and Moore, 2007).

While use of automatic metrics is increasing (Gkatzia and Mahamood, 2015), the above concerns might contribute to the fact that evaluation using human input is still the most popular evaluation method in the NLG domain. Most of these evaluations have been quantitative (Sambaraju et al., 2011), but there can be sizable differences between these quantitative evaluations. A main distinction can be made between intrinsic and extrinsic methods (Hastie and Belz, 2014). Intrinsic methods evaluate the output of the system itself, for example by having humans read and rate text output of the NLG system and comparing these ratings against human written texts on metrics such as fluency, correctness, understandability, etcetera. Extrinsic measures aim to evaluate the impact of the system, for instance by measuring whether a system can fulfill the purpose it was built for (Hastie and Belz, 2014). A corpus analysis by Gkatzia and Mahamood (2015) shows that intrinsic human-based measures are used to a much greater degree compared to extrinsic human-based evaluations, although increasingly more effort regarding the latter evaluation can be observed in the past few years (Gkatzia et al., 2017; Goldstein et al., 2017; Ramos-Soto et al., 2017, for instance)

These extrinsic task-focused evaluations have traditionally been regarded as the type of evaluation

that provides the most meaningful results. However, it can be an expensive and timely undertaking to execute such an evaluation (Reiter and Belz, 2009). Another reason why extrinsic evaluations are not used as often is because in many cases, the purpose of the system itself makes such an evaluation unfit to use. An intrinsic evaluation measuring correctness and grammaticality would often be sufficient for systems that are designed to provide a general audience information in a straightforward manner. However, one of the strengths of data-to-text systems is the fact that they can tailor texts towards specific audiences relatively effortlessly. A notable example of this is the BabyTalk project (Gatt et al., 2009), where separate reports about babies in a Neonatal Intensive Care Unit (NICU) are generated for physicians, nurses and the baby's family based on data in the baby's electronic medical record. Recently, an increasing trend might also be observed in methods that can produce a text without too much human input (e.g. sequence-to-sequence models), which would make the development of systems with this feature much easier. Implementation of a tailoring feature, however, also means that the purpose of the system extends beyond merely informing a general audience. Tailoring is often implemented to increase behavior or attitude (Noar et al., 2007, for an overview), which would warrant a more extensive evaluation to measure the outcomes of the computer-generated texts. Furthermore, the effects of tailoring have mainly been investigated using human-written texts. It might be possible that people respond differently to computer-generated texts, for instance because they have different expectations of texts written by computers compared to human-written texts (Graefe et al., 2016).

2.2 PASS and 'Affective' Natural Language Generation

While an increase in NLG systems with a tailoring component might be expected, one of the few systems in this category is PASS (van der Lee et al., 2017). PASS generates soccer match summaries aimed at fans of one of the teams that participated in the match. Thus, the system can be seen as part of the 'Affective' NLG (ANLG) tradition, which aims to produce texts tailored towards the emotional aspects of the intended reader (Mahamood and Reiter, 2011; Ghosh et al., 2017, for instance). The input data is scraped from Goal.com and contains various types of data related to a soccer match (e.g. date played, goals scored, information about the players). Based on this data, a short Dutch match summary of about 50 to 150 words is produced, which is inspired by the reports of the GoalGetter system (Theune et al., 2001). However, although inspired by previous work, texts by PASS are novel in the sense that the templates have been directly derived from sentences in the MeMo FC corpus (Braun et al., 2016). This corpus contains match reports directly taken from the clubs that participated in the match. This often means that the tone of voice in these reports is emotional, while still maintaining a relatively professional style. Using these reports makes it possible to produce tailored match reports with PASS. This means that the tone of a generated report should appear to be more disappointed or frustrated in case the team of the target audience lost, and more upbeat in case of a win for the team of the target audience (van der Lee et al., 2017, for examples).

Empirical evaluation of ANLG systems is considered challenging (Belz and others, 2003; Mahamood and Reiter, 2011) and very few ANLG systems have been tested properly at this point in time. PASS is no exception to that. The lack of extensive empirical testing of that system makes it difficult to determine how well the system performs in terms of text quality and the effectiveness and importance of its 'emotional' tailoring component. Therefore, a more extensive evaluation is necessary in order to assess the perception of the generated texts and the consequences of tailoring.

2.3 Current challenges

Part of what makes evaluation of ANLG systems challenging is finding the right material to compare the output to. One way to compare the output quality of PASS to human texts would be to use texts in the MeMo FC corpus about the same soccer matches, written for the same target audience as reference texts. However, questions arise whether the texts in the MeMo FC corpus are a suitable equivalent for the texts produced by PASS. The texts in the MeMo FC corpus are usually match reports written by people that watched the full soccer match. Thus, it can be assumed that the writers of MeMo FC corpus texts had much more input data to base their reports on. Furthermore, there are no standard guidelines (e.g. text length, text structure) that all writers for soccer team websites adhere to, meaning that the style and

quality of texts in the MeMo FC corpus can be vastly different depending on the writer and the website it was written for.

The use of automated metrics was also deemed to be a non-viable option for the current study if texts from the MeMo FC corpus serve as reference texts, since sentences from the MeMo FC corpus are used as the basis for the templates in PASS. This would make it likely that much overlap occurs between the generated texts and the reference texts, resulting in high scores on metrics such as BLEU and METEOR. However, these scores would say little about how well these templates blend together to form an enjoyable and coherent text that displays the information in a sensible way (Scott and Moore, 2007).

Finding the right reference texts is not the only challenge when evaluating ANLG systems. The effects of tailoring the generated text towards the emotional needs of the intended audience has not received a lot of attention yet. Most research has focused on the effects of tailoring in the context of persuasive messages. A meta-analysis by Noar et al. (2007) shows that tailoring can improve the attitude towards a message and behavioral outcomes, although a relatively small mean effect size for tailoring interventions was found. In the context of texts that aim to achieve other goals rather than persuade, much less is known about the effects of tailoring. Especially in the NLG domain. One of the notable studies in this domain that attempts to study the effects of tailoring is by Reiter et al. (2003) who studied the effects on smoking behavior that tailored smoking cessation letters had. They did not find significant behavioral differences between people that received a tailored vs. non-tailored letter. However, it might be argued that such behavioral changes are difficult to achieve with only a letter. Furthermore, the tailoring aspect was elaborated by tailoring the arguments to the specific challenges that the reader said to face, which is a different kind of tailoring compared to the tailoring of PASS where the style and diction was tailored to fit with the emotions that the reader experiences. This difference in tailoring could also result in different effects.

A study similar to the current study is that of Wann and Branscombe (1992), who investigated mood changes after reading a (human-written) tailored basketball summary. They found that if participants identified strongly with a team involved in the match and if the text was tailored towards fans of that team, the emotions were the most extreme: participants reported the most positive mood if the team won and the most negative if the team lost. Tailoring did not have a significant effect on mood if the participants did not identify strongly with the team. Mahamood and Reiter (2011) found that affective texts were also preferred to non-affective texts and that emotional appropriateness of texts also correlated with understanding ratings when investigating generated BabyTalk texts. These results support the basic premise of PASS that tailoring of sports reports results in behavioral and attitudinal outcomes that non-tailored texts do not achieve. The current study investigated this more deeply, while also investigating the overall perceived text quality of the system. Furthermore, possible moderators of the effectiveness of computer-generated texts on attitudinal and behavioral responses were investigated by looking into possible differences in preconceptions between human-written and computer-generated texts.

3 Evaluation

As previously noted, the aim of the current paper was to study several components of PASS via two evaluation studies. The first study was a human evaluation of the systems output compared to human-written texts. The second study assessed the impact of PASSs tailoring feature. Both evaluation methods will be described more thoroughly in this section.

3.1 Evaluation of text-related attributes and preconceived notions

In this study, texts generated by PASS were compared to human-written texts regarding the perceived text quality, the evidentness of tailoring, and whether participants can accurately identify the writer to be a human or a computer. Furthermore, preconceived notions about human-written and computer-generated texts were investigated.

3.1.1 Participants

A total of 60 people participated in the study (35 women). All these participants were native Dutch college students and had an average age of 21 years and 9 months.

3.1.2 Design

Participants were asked to rate a total of 5 text-pairs (5 human texts and 5 texts generated by PASS). The human texts were written by 14 journalism and communication students, that together, wrote a total of 22 texts. These writers were given the same match statistics as PASS uses to generate a text and they were instructed to write a soccer match summary of a similar size as PASSs reports based on this data. Furthermore, the writers were asked to imagine that they were writing the reports for the club website of one of the two teams that participated in the match, thus writing one soccer report per match. No writers were involved in the rating task. All these instructions were implemented to get the foundation of the human-written text to be fairly equal to the PASS-generated text. The written soccer matches were about soccer matches played in the Dutch second league in the 2015/2016 season. The teams in this division are generally more obscure teams, which would minimize the chance that people's ratings are affected by their love or hatred for one of the teams involved in the match. Furthermore, the computer-generated texts were about the same Dutch second division matches from 2015/2016 and written for the same target audience as the human-written texts. The participants were randomly assigned to one of two versions where each version contained 5 different text-pairs. Counterbalancing was also applied to reduce order bias: so half of the participants in each version received the text-pairs in opposite order from the other half.

Not only were participants provided with a match report, they were also given the match data so that they were able to rate the completeness and accurateness of the information discussed in the report.

3.1.3 Procedure

The study was conducted using *Qualtrics*: an online platform to design surveys. The procedure for all participants was the same. The experiment started with a written instruction and consent form, after which the experiment started. On every page the participants were provided with match statistics and an accompanied text written by either a human or by PASS. The match data was shown so that participants were able to rate the correctness of the information discussed in the report.

After viewing the match data and reading these match reports, participants were asked to indicate whether they thought the text was written for fans of the home team or the away team, whether they thought the text was written by a human or generated by a computer, and why they thought the text was written by a human or generated by a computer. This last question asked for free-text comments to obtain information about preconceived notions participants have in regards to differences between human-written and computer-generated texts. Previous research has shown that these free-text comments often provide valuable insights about generated texts (Reiter and Belz, 2009). These comments were structured based on the text components of (Callaway and Lester, 2002): style (comments on the overall writing style), grammaticality (comments on the syntactic quality of the text), flow (comments on the fluency of the sentences), diction (comments on word choices), readability (comments on how easy to read the text is), logicality (comments on the aptness of the text structure and if information is correctly represented), detail (comments on the amount of detail in the text), and other (uncategorized comments). Furthermore positive, negative, and preconceptions with unclear valence were distinguished for every category.

The participants judged the quality of each text using seven-point Likert-scales on clarity: how clear and understandable the report is ('While reading, I immediately understood the text), fluency: how fluent and easy to read the report is ('This text is written in proper Dutch, 'This text is easily readable), and correctness: how well the information the report is based on is represented in the report itself ('This report does not include extraneous or incorrect information, 'This report does not omit important information).

3.2 Evaluation of tailoring effects

In this study, the effects of tailoring in a soccer match report on perceived text quality was investigated.

3.2.1 Participants

Native Dutch fans of the three most popular and successful soccer teams in the Netherlands: *Ajax*, *Feyenoord*, and *PSV*, were recruited via *Crowdflower*, these clubs were chosen because previous research has found that success increases fan identification (Wann et al., 1994). This resulted in a total of 171 participants (118 male, average age of 29 years and six months). Most of them were Ajax-fans (99), followed by PSV (55) and Feyenoord (47). Supporters of all three teams identified themselves with the club to a similar degree ($F(2, 168) = 1.77$, $p = .17$). Participants also had different educational backgrounds (72 participants college or university; 99 lower education level).

3.2.2 Design

Participants in this study were asked to read and rate a total of four match reports generated by PASS. The match reports that were presented to the participants was based on the club that they supported. The team they supported was involved in all four instances. A between-subjects design with two conditions was used (text tailored towards the team the participant identifies with, or text tailored towards the team the participant does not identify with).

Participants in all conditions got to see similar match summaries. The reports were based on actual matches played by Ajax, Feyenoord and PSV in the 2015-2016 season. By presenting matches based on actual matches, the aim was to present summaries that are realistic to the participants, but not recent enough for participants to remember the match. Perspective was manipulated by generating two reports with PASS, tailored towards fans of each respective team that participated in the match. The matches shown to participants were chosen randomly from the matches that Ajax, Feyenoord and PSV played in the 2015-2016 season, with the exception of any matches that the team they identified with played against rivaled teams, since these matches could result in more extreme ratings (Cialdini et al., 1995). Similarly, no matches of clubs that have been relegated since the 2015-2016 season were shown to avoid that participants view these matches as unrealistic.

3.2.3 Procedure

Participation of this study was done via an online *Qualtrics* survey. After receiving instructions and filling out a consent form, the participants were asked via multiple choice to indicate whether they had a preference for Ajax, Feyenoord, or PSV. After indicating their preference, fans were asked to read and rate four texts. The texts they received were based on their team preference: all match reports shown involved the preferred team.

After reading a match report, participants were asked to rate the text on 10 7-point semantic differentials based on Maes et al. (1996). Five differentials covered an aesthetic judgment on the text ('uninteresting/interesting', 'detached/appealing', 'distant/inviting', 'boring/engaging', 'impersonal/personal', 'monotonous/varied') and five covered clarity ('difficult/easy', 'complicated/simple', 'unclear/clear', 'complex/clear', 'illogically structured/logically structured', 'cumbrous/concise').

4 Results

4.1 Evaluation of text-related attributes and preconceived notions

4.1.1 Text quality

A two-way repeated-measures ANOVA was executed to investigate an effect of computer-generated vs. human-written texts on perceptions of clarity, fluency and correctness. The results showed main effects for all three text quality components (clarity: $F(1, 59) = 9.448$, $p = .003$; fluency: $F(1, 59) = 8.656$, $p = .005$; correctness: $F(1, 59) = 8.302$, $p = .006$). The participants rated the human-written texts as more clear and fluent compared to its computer-generated counterparts. Conversely, the computer-generated counterparts gave a more precise overview of the information it is trying to convey although scores in both categories were low. These low scores might be due to the fact that most human-written as well as most computer-generated texts did not express all the match data that was shown to raters. While the differences were significant, it should also be noted that the differences were relatively small: for all

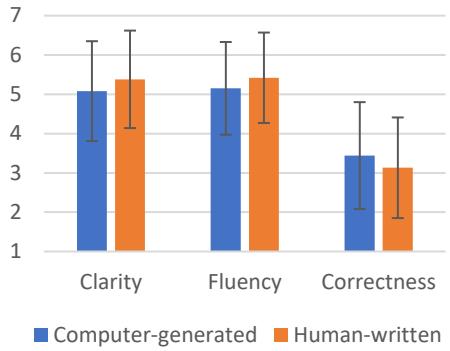

Figure 1: Average scores on clarity, fluency and correctness for the human-written and computer-generated texts on a 7-point Likert-scale. Error bars represent standard deviations.

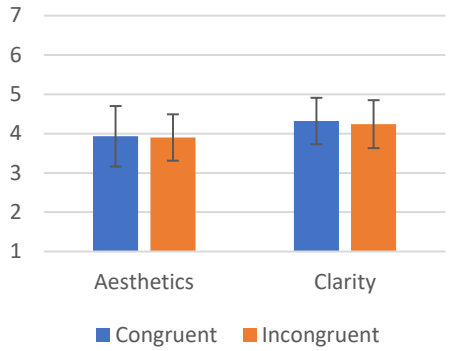

Figure 2: Average scores on aesthetics and clarity for the texts congruent and incongruent with team preference of participants on a 7-point Likert scale. Error bars represent standard deviations.

three text components, the difference between the computer-generated and human-written text was only around 0.3 on a 7-point scale; cf. Figure 1. This would suggest that PASS-generated texts are not that far off the text quality of human texts.

Text type	**Correct perception**	**Incorrect perception**
Computer	246	55
Human	228	72

Table 1: Cross-tab comparing the participants' correct and incorrect tailoring perceptions for computer-generated and human-written texts

Text type	**Computer (perceived)**	**Human (perceived)**
Computer	121	180
Human	94	206

Table 2: Cross-tab comparing the participants' perceived type of text versus the actual text type (computer-generated or human-written)

4.1.2 Perceived tailoring

In general, participants were able to correctly identify the target audience the text was tailored to. In almost 79% of all cases, the participants' perceived target audience was congruent with the intended target audience. An additional chi-square test did not show a significant difference in the evidentness of tailoring between the computer-generated and human-written texts ($\chi^2(1) = 2.96$, $p = .09$). The tailoring was similarly clear for computer-generated texts as for human-written texts; cf. Table 1.

4.1.3 Perceived text type

Results of a chi-square test showed a correlation between the perceived type of writer and the actual type of writer ($\chi^2(1) = 5.14$, $p = .02$). The origin of human-written texts were more often perceived as human and less often as computer compared to computer-generated texts. However, subsequent analysis made it clear that participants had trouble with this task. Mostly for the computer-generated texts: in less than half of the cases were people able to correctly identify the computer-generated text as such (in 40% of cases). Similar to 4.1.1, these results suggest that the texts generated by PASS contains human-like qualities, which was also further investigated in 4.1.4; cf. Table 2.

4.1.4 Free text comments

The free text comments - structured using the aforementioned categories based on (Callaway and Lester, 2002) - revealed clear differences in preconceived notions, most evidently between human-written and computer-generated texts. This was corroborated by a chi-square test for positive and negative notions ($\chi^2(1) = 391.09$, $p < .001$): participants generally had a much more positive stance towards human-written texts compared to computer-generated texts for nearly every text component. People tend to think of human-written texts as more emotional, dynamic, and well-written and computer-generated texts as more static, boring and poorly-written. This was expressed in the style, flow, and readability category. Interestingly, these judgments were not always justified, as 3 and 4 show. Human-written texts

		Positive	Negative	Unclear	Total
	Style	47	0	5	52
	Grammaticality	9	3	2	14
	Flow	8	0	0	8
Incorrectly	Diction	55	0	12	67
Perceived	Readability	10	0	0	10
Human	Logicality	11	9	3	23
	Detail	0	0	6	6
	Other	0	0	5	5
	Total	140	12	33	185
	Style	1	35	6	42
	Grammaticality	0	6	0	6
	Flow	1	3	0	4
Incorrectly	Diction	0	15	7	22
Perceived	Readability	0	2	0	2
Computer	Logicality	4	6	2	12
	Detail	0	10	2	12
	Other	0	0	3	3
	Total	6	77	20	103

Table 3: Frequency of the type of comments in support of participants' incorrectly perceived text type

		Positive	Negative	Unclear	Total
	Style	59	0	5	64
	Grammaticality	8	8	1	17
	Flow	13	0	0	13
Correctly	Diction	62	0	11	73
Perceived	Readability	13	0	0	13
Human	Logicality	16	5	5	26
	Detail	4	2	7	13
	Other	0	0	7	7
	Total	175	15	36	226
	Style	1	53	7	61
	Grammaticality	1	6	1	8
	Flow	0	6	0	6
Correctly	Diction	0	12	5	17
Perceived	Readability	2	4	1	7
Computer	Logicality	4	31	0	35
	Detail	0	4	0	4
	Other	0	0	3	3
	Total	8	116	17	141

Table 4: Frequency of the type of comments in support of participants' correctly perceived text type

were often misjudged as computer-generated, with substantiations such as *less personal*, *boring* and *no emotion*. Similarly, computer-generated texts were mistaken for human-written because they were *vividly written*, *enthusiastically written* and contained *emotions*. Thus, the intended design of PASS, which was to produce texts that are similar to human in style seems to be successful according by these comments.

Further support for this was found in the comments on diction. Participants expected human texts to contain more varied, figurative language and more adjectives (e.g. *use of proverbs*, *lots of variation in word choice*, *use of adjectives*), while computer-generated language was expected to be factual, simple, and sometimes illogical (e.g. *contains lots of numbers*, *illogical word choices*, *simple language use*). Examples of these comments were also found in correct and incorrect attributions of the text type, further suggesting that the computer-generated texts used in this study had similar qualities as human-written texts.

Participants also expected computers to be wrong more often in terms of syntax and information presented. Examples like *contains many errors*, and *grammatically incorrect* were used to explain why the text was perceived as a computer-generated text. These results are especially interesting, because these preconceptions contradict the findings in 4.1.1 where correctness for computer-generated texts was rated (slightly) higher compared to human-written texts.

4.2 Evaluation of tailoring effects

An independent-samples MANOVA was performed on the average scores of the four rated texts. A distinction between aesthetics and clarity was made in the test. For both these components, the MANOVA did not show a significant effect of tailoring congruity (aesthetics: $F(1, 167) = 0.11$, $p = .74$; clarity: $F(1, 167) = 0.61$, $p = .45$). Participants did not find the text to be aesthetically different if the tailoring was congruent with their preference or if it was not. Similarly, participants did not perceive the text as more clear if they were part of the audience tailored to or if they were not; cf. Figure 2. Thus, while the tailoring component was clear, no effects of tailoring were found on attitude towards the text.

5 Discussion

Various aspects of PASS, a data-to-text system that converts soccer match data into a textual soccer match summary, have been evaluated in this paper (van der Lee et al., 2017). PASS is relatively unique in the NLG landscape in the sense that it tailors texts towards specific subgroups. In the case of PASS: supporters of one team or the other involved in a soccer match. This tailoring component is enabled by using texts from the MeMo FC corpus, which contains soccer reports for club websites written in a more emotional tone-of-voice compared to the more neutral newspaper reports. By incorporating a tailoring component, it can be argued that the purpose of PASS is not to merely inform readers of a soccer match, but also to entertain, which required further investigation. Tailoring has been found to increase behavioral and attitudinal outcomes (Noar et al., 2007), and tailored sports reports have been

found to increase emotions for people who identify with a participating team (Wann and Branscombe, 1992). However, if tailoring-related effects could be found for computer-generated texts and if tailoring also affects perceived text quality warranted further investigation. In sum, the aim of the paper was to perform a more extensive evaluation of PASS by comparing its output to similar human-written texts, and to investigate if the attempted tailoring was clear and if it had any attitudinal effects. Furthermore, differences in preconceived notions between human-written and computer-generated texts - a possible moderating factor in the effectiveness of tailoring - was explored. These aims were investigated using two evaluation studies.

The first evaluation study showed differences in perceived quality between human-written and computer-generated soccer match summaries from PASS. The results showed that the language use of the human-written texts was found to be more fluent and easy to read, as well as more clear and understandable. However, the computer-generated texts gave a better overview of the match-data it was based on. The differences in perceived text quality between the human-written and computer-generated texts might be due to the fact that PASS texts are inspired by texts from GoalGetter - a data-to-text system that attempted to produce a neutral and factual match summary (Theune et al., 2001) - in terms of its text structure and the types of data that are incorporated into a text. Perhaps, if the aim of PASS is to entertain rather than inform, it might be fruitful to stray further away from the structure and data use of GoalGetter texts. Interesting, however, was the fact that despite differences in information representation and language use, participants had trouble identifying the computer-generated texts as such: these texts were incorrectly marked as a human text in 60% of cases, which does suggest that the current state of PASS-generated texts is of good quality. Another reason to think that the PASS-generated texts are effective in their goal is the fact that the intended tailoring was correctly identified by participants during evaluations, to a similar degree as the tailoring in human-texts.

However, the second study did not show any effects of tailoring on perceived text quality. Texts that were tailored towards the preferences of the participant were neither seen as more aesthetically pleasing nor as more understandable. Thus, no support was found for tailoring to have an effect on the attitude towards the text. A reason for this might be the findings in the free text comments of the first study. These comments showed that people have negative preconceived notions about computer-generated texts. They are generally expected to be written with boring and unemotional language in a predictable and static style. Knowing that the reports were computer-generated might have biased the participants in the second study to think that the reports were lacking emotions, thus making it harder to achieve tailoring effects. However, it should be noted that this lack of result in the second study does not necessarily mean that there is no effect of tailoring in PASS texts. Previous research has shown that tailored human-written texts about sports matches did increase emotional experiences for participants the texts were tailored towards (Wann and Branscombe, 1992), which could suggest that the PASS texts affect emotional experiences rather than opinions on the text itself. Furthermore, (Mahamood and Reiter, 2011) found that 'emotional appropriateness' correlated with understandability and that 'tailored' texts scored higher on all aspects of text quality compared to 'neutral' texts, which are aspects that are still worth investigating in the context of PASS. There is also a lack of research looking into the relationship between extrinsic evaluation and intrinsic evaluation (Gatt and Belz, 2010), which could be an interesting avenue for further research. Furthermore, the current quantitative human-based approaches could be combined with a qualitative analysis of the output texts similar to McKinlay et al. (2010) and Sambaraju et al. (2011) in order to get a better sense which components of PASS should be improved and how they should be improved.

With the study described in this article, a more extensive evaluation of a data-to-text system was executed than is conventional in the NLG domain, with a few notable exceptions (Reiter and Belz, 2009; Reiter et al., 2003, for instance). We feel that similar evaluations and further research into evaluation methods should be greatly encouraged. Over the last few years, interest in NLG systems has been increasing in domains outside academia as well, such as the journalism domain, and many people expect that computer-generated texts will become more visible in everyday life. In light of these developments, increasing the quality and quantity of evaluations is necessary for a better understanding of the state of these NLG-systems and the role they should play alongside human writers.

Acknowledgements

We received support from RAAK-PRO SIA (2014-01-51PRO) and The Netherlands Organization for Scientific Research (NWO 360-89-050), which is gratefully acknowledged. Thanks are due to Martijn Goudbeek for help in setting up in the experiments, Nadine Braun for providing us with the corpus, Monique Hamers for helping us find young journalists and the three reviewers for their comments.

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.

Anja Belz et al. 2003. And now with feeling: Developments in emotional language generation. *Information Technology Research Institute Technical Report Series*.

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*, pages 313–320. Association for Computational Linguistics.

Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. 2016. The Multilingual Affective Soccer Corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 74–78.

Charles B Callaway and James C Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

Robert B Cialdini, Melanie R Trost, and Jason T Newsom. 1995. Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, 69(2):318.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical Methods in Natural Language Generation*, pages 264–293. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Albert Gatt, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *Ai Communications*, 22(3):153–186.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 634–642.

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. Association for Computational Linguistics.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2017. Data-to-text generation improves decision-making under uncertainty. *IEEE Computational Intelligence Magazine*, 12(3):10–17.

Ayelet Goldstein, Yuval Shahar, Efrat Orenbuch, and Matan J Cohen. 2017. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. *Artificial Intelligence in Medicine*, 82:20–33.

Andreas Graefe, Mario Haim, Bastian Haarmann, and Hans-Bernd Brosius. 2016. Readers perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, pages 1–16.

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

Alfons Maes, Mathilde Maria Nicoline Ummelen, and Hans Hoeken. 1996. *Instructieve teksten. Analyse, ontwerp en evaluatie*. Uitgeverij Coutinho.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21. Association for Computational Linguistics.

Andy McKinlay, Chris McVittie, Ehud Reiter, Yvonne Freer, Cindy Sykes, and Robert Logie. 2010. Design issues for socially intelligent user interfaces. *Methods of Information in Medicine*, 49(04):379–387.

Seth M Noar, Christina N Benac, and Melissa S Harris. 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological bulletin*, 133(4):673.

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.

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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

A Ramos-Soto, J Janeiro, JM Alonso, A Bugarin, and D Berea-Cabaleiro. 2017. Using fuzzy sets in a data-to-text system for business service intelligence. In *Advances in Fuzzy Logic and Technology 2017*, pages 220–231. Springer.

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.

Ehud Reiter, Roma Robertson, and Liesl M Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Rahul Sambaraju, Ehud Reiter, Robert Logie, Andy McKinlay, Chris McVittie, Albert Gatt, and Cindy Sykes. 2011. What is in a text and what does it do: Qualitative evaluations of an NLG system –the BT-Nurse– using content analysis and discourse analysis. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 22–31. Association for Computational Linguistics.

Donia Scott and Johanna Moore. 2007. An NLG evaluation competition? Eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 22–23.

Mariët Theune, Esther Klabbers, Jan-Roelof de Pijper, Emiel Krahmer, and Jan Odijk. 2001. From data to speech: A general approach. *Natural Language Engineering*, 7(1):47–86.

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

Daniel L Wann and Nyla R Branscombe. 1992. Emotional responses to the sports page. *Journal of Sport and Social Issues*, 16(1):49–64.

Daniel L Wann, Thomas J Dolan, Kimberly K MeGeorge, and Julie A Allison. 1994. Relationships between spectator identification and spectators' perceptions of influence, spectators' emotions, and competition outcome. *Journal of Sport and Exercise Psychology*, 16(4):347–364.