HHS Public Access

Author manuscript

Proc Conf Empir Methods Nat Lang Process. Author manuscript; available in PMC 2023 July 07.

Published in final edited form as:

Proc Conf Empir Methods Nat Lang Process. 2021 November ; 2021: 5190–5202.

doi:10.18653/v1/2021.emnlp-main.421.

Refocusing on Relevance: Personalization in NLG

Shiran Dudy,

Department of Computer Science, University of Colorado

Steven Bedrick,

Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University

Bonnie Webber

Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh

Abstract

Many NLG tasks such as summarization, dialogue response, or open domain question answering focus primarily on a source text in order to generate a target response. This standard approach falls short, however, when a user's intent or context of work is not easily recoverable based solely on that source text–a scenario that we argue is more of the rule than the exception. In this work, we argue that NLG systems in general should place a much higher level of emphasis on making use of additional context, and suggest that *relevance* (as used in Information Retrieval) be thought of as a crucial tool for designing user-oriented text-generating tasks. We further discuss possible harms and hazards around such personalization, and argue that value-sensitive design represents a crucial path forward through these challenges.

1 Introduction

The evaluation of natural language generation tasks (such as automatic summarization, machine translation, and dialogue generation) is commonly framed as one of comparing an automated system's generated output to some reference output,[1] with the goal of achieving as close an alignment as possible. Implicit in this experimental framing is the notion that for any given system input, there must exist a single, "correct" output.

While this may arguably be a necessary simplifying assumption in terms of experimental evaluation,[2] this "one-size-fits-all" philosophy has also constrained the ways in which NLG systems are designed, trained, and deployed. For example, standard approaches to automated document summarization and machine translation rely only on the original source text, and do not typically take into consideration contextual factors involving the user or their situation of use. The problem is even larger when we consider NLG systems whose

shdu9019@colorado.edu .

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5190–5202

[1] In some cases, a small set of reference outputs may be used, as in BLEU's original formulation.

[2] Statistical metrics based on this assumption often fail to reflect human judgments of quality or performance; see Novikova et al.'s recent analysis.

outputs are less firmly grounded to some specific input text, such as dialogue and question-answering systems.

This has a number of negative effects on NLG system performance and utility. One such effect is system output that does not meet a given user's needs: for example, a summarization system that chooses to focus on different aspects of the source text than those in which the user was interested, or a machine translation system whose output is of an inappropriate register of formality. Other effects are more subtle; for example, consider that, by designing systems to produce a single universal output, we amplify the effects of label and sample bias in training data (Shah et al., 2020), since when there is only one "right" answer, a statistical model will generally tend towards whatever it has seen most frequently during training. One example of this is the well-documented behavior of machine translation systems defaulting to "male" for certain categories of phrase in gender-marked languages (Prates et al., 2019); another can be seen in the tendency for neural language models to over-predict frequent words (Dudy and Bedrick, 2020).

Recent work by Flek (2020) made a compelling argument for an increased emphasis on user- and task-level personalization in NLP applications, particularly those involved in classification or prediction. In this work, we build on their foundation and turn our attention specifically to tasks involving natural language generation. While the paramount importance of user-level personalization in NLG applications has long been known in our field (Rich, 1979; Kass and Finin, 1988), recent years have seen reduced emphasis on this aspect of system design and evaluation. We challenge the universalist simplifying assumption that underlies much of how such systems are built and evaluated today, and discuss how concepts from information retrieval — specifically different notions of *relevance* — may be fruitful tools for considering personalization in the context of common NLG tasks. We discuss several case studies in which an NLG system would benefit (or indeed, require) additional context about the user and their task, and discuss some possible challenges and harms that could be introduced by textual personalization in NLG. Finally, we discuss how design methodologies that center the needs, goals, and rights of users may provide a crucial path forward to developing more robust and personalized NLG systems; specifically, we point to Value-Sensitive Design (Friedman and Hendry, 2019) as one such option.

1.1 The Universal Truth Assumption

The notion of a ground truth in which each annotated instance in a dataset corresponds to a single "right" answer has been already challenged in the work of Aroyo and Welty (2013, 2015) in the context of language processing, where the authors proposed that instead of the ground truth assumption of a single correct target to a particular task, the crowd truth approach assumption generalizes more optimally as they reflect a set of various perspectives and interpretations, rather than one based on a single answer. Aroyo and Welty argued that in many NLP tasks, low inter-agreement among annotators is expected due to relatively ambiguous instances (sources) which subsequently triggers annotators in different ways, explaining the multiple perspectives (targets) collected.

We argue there is room for more than a single strategy to overcome the ambiguity of the source, so while Aroyo and Welty proposed to embrace the diverse set of responses resulted

from ambiguous input, we propose to further specify the under-specified source in order to generate more relevant outcomes to a particular user or scenario. Particularly, we consider scenarios in which the source may be specified by adding indirect information about the user and their situation, and thereby contribute to reducing the ambiguity to a more limited set of predictions so that, given the same source, the target output may vary depending on the additional meta-data provided.

Following that, one may ask *what* indirect and contextual information would be appropriate for use in textual personalization of NLG systems. Necessarily, there is no single answer to this, as it is entirely task- and user- dependent. Similarly, when we consider the question of *evaluation* of NLG systems, we also find a similar situation: there is no single universal "right" way to evaluate an NLG system (in the way that, say, variations on classification accuracy are generally appropriate for most classification problems). Even "zooming in" on a specific NLG task — abstractive summarization, for example — notions of what makes a "good" summary quickly become worryingly fuzzy,[3] and the situation only becomes more so when we begin to account for issues of user-level personalization.

To move forward, we draw inspiration from the field of information retrieval, which has long faced a similar quandary in its decades-long quest for agreement on the notion of "relevance."

1.2 Relevance: A Path Forward

The key purpose of information retrieval systems has historically been framed as one of providing users with documents (or other information objects, such as images, etc.) that are "relevant" with respect to an information need. The question of how best to define what is meant by "relevant" has dogged the field since its inception (see Saracevic (1975) for a classic early review of the topic[4]), and has been the focus of a great deal of research from many different perspectives. Borlund (2003) grouped this body of work into two broad families of models of relevance. The first, and oldest, frames relevance as a question of topicality: a search result was relevant if it addressed a topic included in a search query. This model has certain similarities to the way that we often think of NLG tasks: the system's behavior is measured in terms of the relationship between a given input and a given output, compared in a fairly constrained way.[5] The user is absent from the conversation, as are specifics of what they are trying to accomplish. This model is sometimes framed as a "system-oriented" or "objective" model of relevance.

This model is not without its uses, but has important limitations. As a simple illustration of this, consider that a given search result may be topically aligned with some aspect of a search query… but if it covers information our user already knows, can it truly count as being "relevant?" The second family of relevance models, labeled by Borlund and others as "user-oriented" or "subjective" relevance, addresses these limitations by conceiving of "relevance" as a more complex, dynamic, and multidimensional construct. Crucially, this

[3]See section 2.2 for further discussion on this point.
[4]See Saracevic (2006, 2007) for more recent reviews.
[5]In IR, the question is whether or not a concept from the query is mentioned in the document; where analogously in NLG we use overlap-based metrics such as BLEU or ROUGE.

view of relevance necessitates the consideration of who, exactly, we are imagining as our user, and what it is that they are attempting to accomplish.

There are many different models of user-oriented relevance, but for the purposes of considering NLG applications, we find *situational relevance* (Schamber et al., 1990) to be particularly useful. Under this model, relevance is understood as "the utility or *usefulness* of the … information objects" in terms of "the relationship between such … objects and the work task at hand underlying the information need as perceived by the user" (Borlund, 2003, p. 915, emphasis original).[6] Another related notion of relevance that is highly applicable to NLG is that of task-oriented relevance, in which the notion of relevance regarding a system output is explicitly tied to that output's impact on the user's ability to complete their task. As defined by Hjørland and Christensen (2002), "Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T."

While there are many types of NLG applications, with many different objectives, all of them share the underlying goal of communicating a given text/utterance to a human user in helpful and relevant way. Looking through the lens of situational and task-oriented relevance, we posit that for any such application, the notion of there being a single "correct" output for a given input is nonsensical. Adopting a one-size-fits-all approach to system development and evaluation is thus more than a simplifying assumption; it ignores a key and integral aspect of system behavior.

Only by taking into account the user and their task can the output of a system be made relevant; one way to achieve that is through communicating information to a user in ways that are relevant to them and/or their situation. Early work on personalized natural language systems such as that of Kass and Finin (1988) recognized this, and placed heavy emphasis on building systems around rich user models. More recently, work by Newman et al. (2020) makes the link directly to the core purpose of an NLG system, framing the problem as one of modeling the "communicative function of language." They point out that "a speaker's goal is not only to produce well-formed expressions, but to convey relevant information to a listener"; in the context of NLG, this must necessarily take a personalized form.

As an example of how a user's immediate situation may relate to the content and pragmatics of generated language, consider a user who asks a question-answering system "how to put out a fire." If they are in the kitchen standing over a grease fire, a situationally-relevant response would be short in length, very focused on a specific firefighting technique, and delivered in an imperative register. In a different situation (perhaps while sitting around the dinner table with inquisitive children), a situationally-relevant response to the same query might include more generic informational content about firefighting, and be delivered in a more narrative mode.

[6] *Situational relevance* as introduced by Schamber et al. is meant as a broad and somewhat abstract term, and encompasses to the totality of the situation surrounding the use of a tool, including the user, their setting, and their task. Elsewhere, the word "situation" is sometimes used in a more narrow and concrete manner, referring to either the setting in which a task takes place or to a specific task itself.

Following this example, we now turn discuss several different NLG application areas, and the role personalization may play in each.

2 A Survey of Textual Personalization

2.1 The Status Quo

The goal of textual personalization in NLG is to develop models that generate *relevant* text to users in response to information (or other) needs, expressed in natural language, the exact form of which will depend on the specific application. Asking a question, or interacting in a dialogue, does not happen in a void; rather, there is an underlying intent for doing so. The problem is that current text generation tasks operate independently from the users they aim to support, which subsequently limits their usefulness when deployed. What are the limitations introduced by not modeling users in the system?

In abstractive summarization the dominant assumption is generally that there is a single type of summary that can be produced from a given text. This applies also to evaluation scenarios that make use of multiple reference outputs, as in such cases an implicit assumption is that the references should typically exhibit minor linguistic variation, rather than summaries that vary substantially in their contents (Cachola et al., 2020; See et al., 2017; Grusky et al., 2018; Harman and Over, 2004). While this approach simplifies the development and evaluation process, in practice, different users would likely find different aspects of the source article to be more relevant to their needs than others; in other words, if an original article includes facts *A* through *E*, one user's optimal summary might involve facts *{A, B, C}* while another user's would instead feature *{A, C, D}*. This example will be referred as the source-to-target transformation example. Beyond the factual content of a summary, users could also vary in terms of the level of detail that they would find useful (Louis and Nenkova, 2011), some would enjoy in-depth summarization, while other users would benefit from a simplified writing style (Scarton et al., 2018a). Similarly, in a paraphrasing (Witteveen et al., 2019) task, two different users, each with different intents and goals, would likely find different paraphrases to be "correct." A development and evaluation paradigm which assumes a single reference output (or a small set of semantically-equivalent reference outputs), is unlikely to support (or encourage) the generation of user- and task-personalized output.

Thus far, we have described families of task that we will henceforth refer to as *source-to-target transformation* generation tasks. This category encompasses tasks that are firmly grounded in a specific source text, and which must produce fluent textual output whose content is similarly grounded to that same source, e.g. automated summarization, paraphrasing, narrative report generation, etc. We will now discuss a second category of generation, which we refer to as *query-response*. and includes question-answering, dialogue agents, and the like. Query-response tasks are generally guided by a query or prompt from a user (or by a series of such prompts), and while they may draw on source documents of various sorts to inform their outputs, the link is much less straightforward than in the transformation tasks that we have previously discussed. User-level personalization, however, is perhaps even more essential to query-response tasks. Consider, for example, open domain question answering: for any given question, there might be multiple factually-correct

answers (Yang et al., 2015), but without modeling a user in this dynamic, it is difficult to say whether any particular answer will or will not be be relevant.

Open domain dialogue systems (Li et al., 2016b; Wen et al.) face a somewhat related problem to what is described above from a technical stand-point, in that there is no one ground truth that is expected. Dialogue agents tend to generate general prompts that may address a given question (the query) and present a natural and informative response, but are indifferent to the user, irrespective of anything but the question it was asked about. Crucially, however, there has been significant attention paid to the problem of personalizing dialogue agents, to a much greater degree than is the case in other realms of NLG.

Some of this work has focused on the use of psycho-linguistically informed parameters (verbosity, etc.) to tune a system's output (Mairesse and Walker, 2011), while other work has focused on ways to make use of more general "personas," meant to represent salient features of the dialogue agent or its interlocutor (or both), and to use those features to influence the semantic and stylistic content produced by the agent (Li et al., 2016a; Zhang et al., 2018). Going the other direction, Madotto et al. (2019) and others have worked to infer relevant properties of the interlocutor from the conversation itself, rather than relying on a pre-specified persona. An additional direction of work in dialogue personalization has been in efforts to have automated dialogue agents behave "empathetically" with their interlocutors, by attempting to match the register and contents of their output with what they perceive to be the emotional state of their user (see e.g. Lin et al., 2019).[7]

It is perhaps unsurprising that dialogue systems have focused on personalization to a larger extent than have other types of NLG application, as dialogue systems are deeply and necessarily user-oriented, both in terms of their design and their evaluation, in a way that other types of NLG are not traditionally thought of as being. In human-human interaction we communicate in a more collaborative fashion, considering what may be additional information required to solve a problem, knowing that the same question might be responded differently depending on the user's age, situation, gender, expertise, register, patience, and underlying intent when posing a question.

Thus, when a dialogue system *fails* to perform in this way, it represents an obvious failure of the system, much more than slightly agrammatical output in a generated summary might, for example. Put in terms of the "fluency" and "adequacy" dimensions often used in MT evaluation (White et al., 1994), the two are much more closely tied together in the case of a dialogue system than they are in an MT system. All of that said, current neural-network-based dialogue systems are very much in their infancy with regards to personalization; in a recent work discussing aspects of human-computer interaction in relation to dialogue systems, Kopp and Krämer (2021) suggest that the field should "(re-)emphasize the hallmarks of human communication and its complexity, and … argue that we should not lose sight of these hallmarks when deriving requirements for human-agent-interaction." In our work, we are particularly interested in shedding light on the end user and their needs when using textual communication systems.

[7]These directions can be combined (Zhong et al., 2020).

2.2 Benefits of Personalization in NLG

Different NLG tasks can benefit from personalization in different ways. For example, the utility of an automatically-generated document summary will depend heavily on the context of use, as noted by Sparck Jones (1998), who listed several factors that come into play in the task of human summarization, specifically in terms of a *purpose* factor that considers the situation, audience, and use – all of which are context dependent. As an example, consider a scientific article that describes a dataset, a model architecture, and an evaluation methodology. A summarization method that takes into account the needs and interests of its user may be able to tailor its output to focus on the most contextually-relevant aspect of the article. Similarly, paraphrasing systems could benefit from a notion of situational relevance, as different users will need different aspects of the source document to be retained, depending on their scenario of use.

Context is also critical for open-domain question-answering, as many questions are under-specified when considered solely in terms of their textual contents. Consider the question "how to drill a hole in the wall?" Without knowing more about the tools available, the composition of the wall, etc., this question is simply unanswerable; a one-size-fits-all attempt is likely to be irrelevant at best.

In terms of *dialogue response*, Adiwardana et al. (2020) presented a system that allows for diverse responses to be elicited by the same prompt, and ? proposed steps towards measuring content diversity of responses, which is aligned with the crowd truth proposal mentioned earlier (Aroyo and Welty, 2013). While this diversity should be encouraged and expected, the next step is to learn how to generate these responses in a more *controllable* way that could enable us to reason why the particular target was generated, instead of consenting to randomly plausible response. One way to get this control is by constructing responses that are relevant in a particular context, either providing them as additional information on the user or situation, or training agents to identify and resolve knowledge ambiguities through clarifying questions in order to respond to a user in a relevant way.

In addition, perceiving the diverse responses to be a result of *additional factors* other than the source alone contributes to a more coherent and explainable outcome enabling measurement of the effect of the additional data on the final target.

For instance, in a scenario where a customer speaks to a hospital representative, the same question might elicit one response if the caller is a patient, and a very different kind of response (in terms of both content and register) if the caller is a medical provider. So while there may be various plausible/correct and diverse responses generated by the system, some of them are mutually exclusive and are dependent on the additional data provided. Furthermore, in this scenario, we note that some possible responses may bear a risk of compromising sensitive data. As another example, a question that a child raises to a voice assistant such as Alexa is expected to be answered differently than when raised by an adult, often in a simplified fashion as described by Scarton et al. (2018a); this task may also involve revising both the content and the syntax of the response.

3 Potential Harms from Textual Personalization

Potential harms from personalization can be grouped into several categories. The first stems from the fact that personalized information systems must necessarily have access to and make use of personal information about their users, which brings with it the risk of that private data being improperly disclosed or otherwise misused (Krishnamurthy et al., 2011; Corrigan et al., 2014). A second, related category of potential harm is that of such an information system being used by its operators to leverage its capabilities against the interests of its users; possible scenarios include targeted advertisements for predatory educational or financial products being shown to vulnerable individuals, or a system that shows or hides certain job listings based on the gender of its user (or, in truth, based on the system's limited model thereof).

An additional, more subtle family of risk has to do with the indirect consequences of widespread deployment of personalized information systems. In social media, and search engines, the choices made by the model developer when incorporating the users' data are based on the confirmation bias theory proposed by Nickerson (1998), describing the tendency to favor information that confirms one's prior beliefs.[8] The assumption is that presenting a user with evidence that supports their belief would increase their engagement, resulting in higher margins for these companies. Strategizing around this goal was shown to contribute to the formation of 'filter bubbles'/'echo chambers' decreasing the users' exposure to diversity of perspectives (Flaxman et al., 2016; Chitra and Musco, 2020) as well as radicalization (Maddox and Creech, 2020). This intellectual isolation can erode the healthy functioning of democracies, which thrives on exposure to multiple points of view (Downs et al., 1957; Baron, 1994; Lassen, 2005).

To illustrate this, Golebiewski and Boyd (2018) describe the phenomenon of searching on "fragmented concepts," where politically-involved queries return only results that support the query's narrative instead of exposing the user to both sides. Golebiewski and Boyd argue that intentionally avoiding returning the results containing the other side stems from the fear of having the search engine portrayed as politically biased. Users are often unaware of the active role that personalized software is taking in this process; Tripodi (2018) describes an interview with a participant who "admitted that her Google searches rarely revealed alternative points of view. However, she did not consider how her returns were tied to her own search practices or Google's algorithmic ordering of information." Tripodi's interviews reveal how users ascribe the lack of alternative views to their nonexistence, rather than search manipulation.

Finally, another possible concern in the process of personalizing a system is that the system's assumptions (and thus its personalized model of the user) may be incorrect, resulting in output that is irrelevant at best, and misleading at worst. For instance, a recipe system supporting a user in the task of making enchiladas might instruct its user to "fold in the cheese"; depending on the user's level of experience as a cook, this may be an

[8] As opposed to vulnerabilities related to a specific user, confirmation bias susceptibility is across users and therefore is mentioned separately.

unfamiliar usage of the word “fold,” and it may be more appropriate to instruct them to “gently incorporate the cheese with a scooping-and-folding motion” (Christensen, 2013). Similar issues could arise in other contexts, with more serious results; imagine a system providing personalized medical advice, for example.

4 Value-Sensitive Design

To design a system that can address the shortcomings described in Section 3, we can follow research in value-sensitive design (VSD, Friedman and Hendry, 2019) that is aimed at identifying the stakeholders and prioritize the values to instill in the system under the assumption that “in designing tools we are designing ways of being.” How might this affect the direct or indirect stakeholders of the system? In this section, we focus on the goals of the user of the system. There is of course no single answer to what this looks like, but in this section we attempt to dissect the problem in order to evaluate the possible outcomes of personalization through the lens offered through VSD.

One value achieved by personalization is that of accessibility of information. A summary written for a child may be realized differently than one for an adult. By learning about the user, the system can adjust the syntax/register/style it produces, as well as the content, to make it more comprehensible to the user (Scarton et al., 2018b). However, this requires that privacy also be taken into account, as personalized systems make use of data about their user. Since these values may be at odds with each other, understanding their possible trade-offs is necessary. By cautiously balancing the most relevant data the system requires for personalization, as well as the privacy concern of its users, system designers may optimize the amount and types of data that the user may be required to turn in.

Another indirect outcome of accessibility of information can be shown in discourse generation, where, for example, the quality of the interaction can elicit user responses that can more easily (or more quickly, or more accurately) direct the system towards the user’s goals of the conversation. In other words, an accurate and thorough (language) understanding of the user’s response may help build a shared ground and ease the transfer of communicating ideas, intents, or requests which subsequently accomplish the user’s goal in a conversation.

Safety can be another important value to be accountable for as developing models that interact with humans. Dinan et al. (2021) elaborate how a dialogue agent can cause real world harm by proactively introducing harmful language or content,[9] or posing as an expert, which can mislead and create harm in the physical world (providing medical advice for COVID-19, for instance). Of course, source-to-target transformation tasks are susceptible to similar problems, as well.

Another opportunity such systems bear is that of enabling a proactive personalization that can provide a sense of agency to a user. This can be done through allowing them to define their specific needs and interests, and by that providing a sense of control over the system as

[9]What Dinan et al. refer to as “instigating.”

described by Synofzik et al. (2008). A personalized system can find ways to reason/explain on the type of data presented to its users, promoting transparency. For instance to paraphrase a sentence for a user, a system can indicate what were the relevant dimensions it maintained and why (see the source-to-target example in section 2.1). Moreover, if a user knows how their data is used that may contribute to developing trust. Ribeiro et al. (2016) identified two types of trust; the first, whether a user trusts an individual *prediction* sufficiently to take some action based on it, and the second, whether the user trusts a *model* to behave in reasonable ways if deployed "in the wild." We add to that a third sense of trust, can we trust the *developer*, that their knowledge integrated into these systems is serving the needs of the user faithfully and respectfully. Projecting aspects of value-sensitive design into text personalization tasks is the first step towards making a human-centered communication system.

5 Architectures for Personalization

The previous section described the conceptual guidelines we propose, while this section is focused on practical aspects on how to implement a language generation system that allows for textual personalization to a situation or a user. At the very basic level, personalization of text can be done through keywords, much as a search engine attempts to return relevant text to a user in response to a query. However, keyword retrieval may not always recover the user's intent and as a result, the returns may not always be relevant. For instance the concept of "neural network" is highly associated with a particular sense, but has another sense that is actively used in the field of neuroscience, making some matches irrelevant. In this work we claim that the advances of natural language generation should be incorporated to personalization in ways that could benefit the user beyond good search engine returns.

Since many NLG tasks are based on language models, we propose architectures that can generate personalized text through the personalization of language models. Khalifa et al. (2021) present in their work a language generating system that enables generating text based on pre-trained language models, that can both apply pointwise as well as distributional constraints, while deviating a little as possible from the original distribution. The essence of this work in the context of personalization is the ability to generate text in a *controlled* fashion. The aspect of control is crucial for the development of future systems. One could adapt a language model to a particular user through conditioning on all outputs to speak about topics relevant to the user (a pointwise constraint) and have 50% of them to be related to topic discussed in the given query (a distributional constraint).[10] A possible concern, at this point, when employing this model has to do with the ability within the same model to adapt different distributional constraints quickly.

Another direction that could overcome this limitation presented by Li and Liang (2021), where the model is conditioned on a prefix intended to provide context to steer the generated text. One way to incorporate this model is to condition the textual generation on different users, however, compromising centralized models bears higher costs to overall users' privacy (Zyskind et al., 2015); instead, a per-user model can be trained on user's

[10]The original example conditioned all predictions to relate to sports, and 50% of them reflect female characters.

personal data, perhaps remaining entirely on their device. Moreover, ideally, implementing personalized agents can be done through Federated Learning Li et al. (2020), where the model updates are made locally and globally through maintaining the user's privacy.[11] As an example of applying the prefix scenario (by Li and Liang (2021)) to personalization, providing a prefix of "How does the email my mom sent relate to my sister?", together with the email in question as the source, would then generate a personalised textual output.

While the focus of this work is mainly on text, it is plausible to assume that the greater objective of such work is to improve and enhance these communication models through the process extending the shared knowledge both parties maintain for the purpose of problem solving. Therefore, personalization may likely be more effective through integrating modalities beyond text and should be a cross disciplinary effort. When considering architectural perspectives for textual personalization, the realization of the information can vary greatly both across different individuals and within the same individual. In their work, following a survey on communication preferences in adults, Himmelsbach et al. (2015) concluded the due to high variation, it is generally recommended to support several modalities of communication, showing that there is no one size fits all approach.

In a survey on multimodality in educational setting, Walters (2010) describes the importance of the impairment-specific approaches to accommodate individuals with specific disabilities and contributing to an inclusive environment; furthermore, Takayoshi and Selfe describes how "the more channels students (…) have to select from when composing and exchanging meaning, the more resources they have at their disposal for being successful communicators." The optimal modality may also depend on the situation, and it is reasonable to consider an event in which a user may opt for more than one modality within the same task; in the example of a driver asking Siri "how can one recognize Mount Saint Helens?", while an image would be a highly informative modality, it would provide a less optimal output at that precise point in time. Instead, in order to avoid removing the driver's sight from the road, a verbal description would be a less risky alternative. At times one modality may complement another as shown by the work of Wang et al. (2016) and Zhu et al. (2018), in which the output of a summarization task was presented not only through text but through images; in Wang et al. (2016) the output was also structured temporally on a timeline. Liao et al. (2018) introduced a multimodal dialogue agent for fashion retail where the visual appearance of clothes and matching styles are crucial in understanding the user's intention. Returning to an earlier example, multimodality can help overcome a system's limitations. Recall the example of a recipe system providing its beginner-level user with the (unhelpful) instruction to "fold in the cheese;" if the instruction was accompanied by a video demonstrating of the process, this could help the user not only accomplish their immediate goal but also to expand their capabilities for next time by learning new vocabulary.

6 Conclusions: How to Begin?

One major obstacle to the development of personalized NLG systems is that such systems often depend on access to sensitive and personal data from users, making large-scale data

[11] In this case, protecting privacy comes at the cost of increased complexity.

resources difficult (or impossible) to obtain and share across a community. While this is a major challenge, we can draw inspiration from related fields that must work with such data and have developed best practices around how to do so, such as the biomedical domain. One possible model is that of the MIMIC-III (Pollard et al., 2016) dataset, which is made up of thoroughly deidentified electronic medical records from thousands of patients from Beth Israel Deaconess Medical Center in Boston, MA. This dataset is freely available to community members to use in research, but access is closely managed. Before gaining access to the dataset, users are required to take an online human-subjects research training module, which also contains content on the protocols for working with the dataset. Users of MIMIC-III must also sign a data use agreement under which they (and, importantly, their institutions) legally agree to restrictions on the use and redistribution of the data.

This solution has proven adequate for the MIMIC-III dataset; however, there exist datasets where even heavily controlled redistribution is not an option. In such cases, *data enclaves* can be effective tools for community research on large, private datasets (Lane and Schur, 2010). One example of such an enclave is the National COVID Cohort Collaborative (N3C) system (Haendel et al., 2020), which is run by the National Institutes of Health and stores harmonized demographic and clinical data about millions of patients from participating health care organizations. After obtaining approval by their local institutional review board (IRB) and executing a formal data use agreement, users who have been granted access are able to analyze and interact with the dataset in a secure cloud computing environment. The environment is designed such that data may not be removed or exported, and the platform includes capabilities for machine learning, statistical analysis, and data visualization. A similar enclave was used during the 2021 CL-Psych Shared Task (MacAvaney et al., 2021) in an NLP context, in order to allow the research community to interact with sensitive data relating to mental health in an IRB-controlled manner.

A related approach has been used in information retrieval in situations where even limited and controlled access to the original data is not possible. Under the “Evaluation-as-a-Service” paradigm (Lin and Efron, 2013; Eggel et al., 2018), developers virtualize their systems, and send them to a secure computing environment where they are trained and evaluated on an entirely private dataset, with the results being shared among the community. While more limiting than a data enclave approach, experience in IR has shown that this model is a feasible one for shared-task evaluation (Hopfgartner et al., 2018). Such a system could be constructed for the design and evaluation of NLG applications involving sensitive data about individuals.

It may also be possible to augment more traditional data sets to simulate personalized behavior without actually requiring sensitive data. For example, in the context of a paraphrasing task, one might supplement the original passage to be paraphrased with a set of features or questions that the algorithm should use to ground its behavior; the same input passage could be repeated with different grounding questions, and the evaluation design could take into account the degree to which the system’s output was responsive to the additional data.

Similarly, in a summarization context, one could augment an input document with different sets of questions representing different users' needs, and an evaluation could take into consideration how well the output summary addressed the stated questions. Open domain questions-answering datasets and tasks could be similarly constructed. In dialogue systems, beyond the persona-based approach previously discussed, there is a long tradition of explicit user modeling (Biswas and Robinson, 2010; Walker et al., 2004; Kass and Finin, 1988) with many approaches that could be drawn on in order to generate simulated needs or goals for such a system to meet.

Another useful avenue to explore is that of richer evaluation methodologies. Rather than restricting our analysis to the system's output, truly assessing whether an NLG system is producing relevant personalized output necessitates extrinsic, task-based evaluations involving users. Such methods have long been used in the evaluation of automated summarization systems (Hand, 1997; He et al., 1999; Mani, 2001; McKeown et al., 2005) as well as NLG systems (Mellish and Dale, 1998; Reiter et al., 2001; Colineau et al., 2002), though we note that recent years have seen somewhat less of this sort of ecologically valid evaluation, and much more focus on statistical evaluation; the work of Barker et al. (2016) and Newman et al. (2020) represent examples of very welcome exceptions to this trend.

7 Future work

In this opinion paper we argue for the benefit of personalizing NLG tasks. We hope that through this work and others' we will continue to make steps towards personalized text, from developing relevant focused datasets through methods that would make text more accessible and of practical use for real users.

Acknowledgements

We would like to thank the anonymous reviewers who provided very useful comments. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805, as well as the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC015999. The opinions expressed are those of the authors, and do not represent views of the NSF or NIH.

References

Adiwardana Daniel, Luong Minh-Thang, So David R, Hall Jamie, Fiedel Noah, Thoppilan Romal, Yang Zi, Kulshreshtha Apoorv, Nemade Gaurav, Lu Yifeng, et al. 2020. Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

Aroyo Lora and Welty Chris. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM, 2013(2013).

Aroyo Lora and Welty Chris. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine, 36(1):15–24.

Barker Emma, Paramita Monica, Funk Adam, Kurtic Emina, Aker Ahmet, Foster Jonathan, Hepple Mark, and Gaizauskas Robert. 2016. What's the Issue Here?: Task-based Evaluation of Reader Comment Summarization Systems. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3094–3101, Portorož, Slovenia. European Language Resources Association (ELRA).

Baron David P. 1994. Electoral competition with informed and uniformed voters. American Political Science Review, pages 33–47.

Biswas Pradipta and Robinson Peter. 2010. A brief survey on user modelling in hci. In Proc. of the International Conference on Intelligent Human Computer Interaction (IHCI), volume 2010.

Borlund Pia. 2003. The concept of relevance in IR. Journal of the American Society for Information Science and Technology (JASIST), 54(10):913–925.

Cachola Isabel, Lo Kyle, Cohan Arman, and Weld Daniel. 2020. TLDR: Extreme summarization of scientific documents. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4766–4777, Online. Association for Computational Linguistics.

Chitra Uthsav and Musco Christopher. 2020. Analyzing the impact of filter bubbles on social network polarization. In Proceedings of the 13th International Conference on Web Search and Data Mining, pages 115–123.

Christensen Emma. How to fold egg whites or whipped cream into a batter [online]. 2013.

Colineau Nathalie, Paris Cecile, and Vander Linden Keith. 2002. An Evaluation of Procedural Instructional Text. In Proceedings of the International Natural Language Generation Conference, pages 128–135, Harriman, New York, USA. Association for Computational Linguistics.

Corrigan Hope B, Craciun Georgiana, and Powell Allison M. 2014. How does target know so much about its customers? utilizing customer analytics to make marketing decisions. Marketing Education Review, 24(2):159–166.

Dinan Emily, Abercrombie Gavin, Bergman A Stevie, Spruit Shannon, Hovy Dirk, Boureau Y-Lan, and Rieser Verena. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. arXiv preprint arXiv:2107.03451.

Downs Anthony et al. 1957. An economic theory of democracy.

Dudy Shiran and Bedrick Steven. 2020. Are some words worth more than others? In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pages 131–142, Online. Association for Computational Linguistics.

Eggel Ivan, Schaer Roger, and Müller Henning. 2018. Distributed container-based evaluation platform for private/large datasets. In 2018 17th International Symposium on Parallel and Distributed Computing (ISPDC), pages 93–100.

Flaxman Seth, Goel Sharad, and Rao Justin M. 2016. Filter bubbles, echo chambers, and online news consumption. Public opinion quarterly, 80(S1):298–320.

Flek Lucie. 2020. Returning the N to NLP: Towards contextually personalized classification models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7828–7838, Online. Association for Computational Linguistics.

Friedman Batya and Hendry David G. 2019. Value sensitive design: Shaping technology with moral imagination. MIT Press.

Golebiewski Michael and Boyd Danah. 2018. Data voids: Where missing data can easily be exploited.

Grusky Max, Naaman Mor, and Artzi Yoav. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 708–719.

Haendel Melissa A, Chute Christopher G, Bennett Tellen D, Eichmann David A, Guinney Justin, Kibbe Warren A, Payne Philip RO, Pfaff Emily R, Robinson Peter N, Saltz Joel H, et al. 2020. The national covid cohort collaborative (n3c): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association.

Hand Thérèse Firmin. 1997. A Proposal for Task-based Evaluation of Text Summarization Systems. In Intelligent Scalable Text Summarization.

Harman Donna and Over Paul. 2004. The effects of human variation in duc summarization evaluation. In Text Summarization Branches Out, pages 10–17.

He Liwei, Sanocki Elizabeth, Gupta Anoop, and Grudin Jonathan. 1999. Auto-Summarization of Audio-Video Presentations. In Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pages 489–498, New York, NY, USA. Association for Computing Machinery.

Himmelsbach Julia, Garschall Markus, Egger Sebastian, Steffek Susanne, and Tscheligi Manfred. 2015. Enabling accessibility through multimodality? interaction modality choices of older adults. In Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia, pages 195–199.

Hjørland Birger and Christensen Frank Sejer. 2002. Work tasks and socio-cognitive relevance: A specific example. Journal of the American Society for Information Science and Technology, 53(11):960–965.

Hopfgartner Frank, Hanbury Allan, Müller Henning, Eggel Ivan, Balog Krisztian, Brodt Torben, Cormack Gordon V., Lin Jimmy, Kalpathy-Cramer Jayashree, Kando Noriko, Kato Makoto P., Krithara Anastasia, Gollub Tim, Potthast Martin, Viegas Evelyne, and Mercer Simon. 2018. Evaluation-as-a-service for the computational sciences: Overview and outlook. J. Data and Information Quality, 10(4).

Kass Robert and Finin Tim. 1988. Modeling the user in natural language systems. Computational Linguistics, 14(3):5–22.

Khalifa Muhammad, Elsahar Hady, and Dymetman Marc. 2021. A distributional approach to controlled text generation. In International Conference on Learning Representations.

Kopp Stefan and Krämer Nicole. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. Frontiers in Psychology, 12:597.

Krishnamurthy Balachander, Naryshkin Konstantin, and Wills Craig. 2011. Privacy leakage vs. protection measures: the growing disconnect. In Proceedings of the Web 2.0 Security and Privacy Workshop.

Lane Julia and Schur Claudia. 2010. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health services research, 45(5p2):1456–1467. [PubMed: 21054366]

Lassen David Dreyer. 2005. The effect of information on voter turnout: Evidence from a natural experiment. American Journal of political science, 49(1):103–118.

Li Jiwei, Galley Michel, Brockett Chris, Spithourakis Georgios, Gao Jianfeng, and Dolan Bill. 2016a. A persona-based neural conversation model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Li Jiwei, Monroe Will, Ritter Alan, Jurafsky Dan, Galley Michel, and Gao Jianfeng. 2016b. Deep reinforcement learning for dialogue generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202.

Li Tian, Sahu Anit Kumar, Talwalkar Ameet, and Smith Virginia. 2020. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3):50–60.

Li Xiang Lisa and Liang Percy. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597, Online. Association for Computational Linguistics.

Liao Lizi, Ma Yunshan, He Xiangnan, Hong Richang, and Chua Tat-seng. 2018. Knowledge-aware multimodal dialogue systems. In Proceedings of the 26th ACM international conference on Multimedia, pages 801–809.

Lin Jimmy and Efron Miles. 2013. Evaluation as a service for information retrieval. SIGIR Forum, 47(2):8–14.

Lin Zhaojiang, Madotto Andrea, Shin Jamin, Xu Peng, and Fung Pascale. 2019. MoEL: Mixture of empathetic listeners. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Louis Annie and Nenkova Ani. 2011. Text specificity and impact on quality of news summaries. In Proceedings of the Workshop on Monolingual Text-To-Text Generation, pages 34–42, Portland, Oregon. Association for Computational Linguistics.

MacAvaney Sean, Mittu Anjali, Coppersmith Glen, Leintz Jeff, and Resnik Philip. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pages 70–80, Online. Association for Computational Linguistics.

Maddox Jessica and Creech Brian. 2020. Interrogating lefttube: Contrapoints and the possibilities of critical media praxis on youtube. Television & New Media, page 1527476420953549.

Madotto Andrea, Lin Zhaojiang, Wu Chien-Sheng, and Fung Pascale. 2019. Personalizing dialogue agents via meta-learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Mairesse François and Walker Marilyn A.. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. Computational Linguistics, 37(3):455–488.

Mani Inderjeet. 2001. Summarization Evaluation: An Overview. In Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. National Institute of Informatics.

McKeown Kathleen, Passonneau Rebecca J, Elson David K, Nenkova Ani, and Hirschberg Julia. 2005. Do Summaries Help? In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 210–217, New York, NY, USA. Association for Computing Machinery.

Mellish C and Dale R. 1998. Evaluation in the context of natural language generation. Computer Speech & Language, 12(4):349–373.

Newman Benjamin, Cohn-Gordon Reuben, and Potts Christopher. 2020. Communication-based Evaluation for Natural Language Generation. In Proceedings of the Society for Computation in Linguistics 2020, pages 116–126, New York, New York. Association for Computational Linguistics.

Nickerson Raymond S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2):175–220.

Novikova Jekaterina, Dušek Ondřej, Curry Amanda Cercas, and Rieser Verena. 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.

Pollard Tom Jay, Shen Lu, Lehman Li-wei, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, and Mark Roger G. 2016. MIMIC-III, a freely accessible critical care database. johnson aew.

Prates Marcelo O R, Avelar Pedro H, and Lamb Luís C. 2019. Assessing gender bias in machine translation: a case study with Google Translate. Neural Computing and Applications, 14(1):1–19.

Reiter Ehud, Robertson Roma, Lennox A Scott, and Osman Liesl. 2001. Using a Randomised Controlled Clinical Trial to Evaluate an NLG System. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 442–449, Toulouse, France. Association for Computational Linguistics.

Ribeiro Marco Tulio, Singh Sameer, and Guestrin Carlos. 2016. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144.

Rich Elaine. 1979. User modeling via stereotypes. Cognitive Science, 3(4):329–354.

Saracevic Tefko. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6):321–343.

Saracevic Tefko. 2006. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II. Advances in Librarianship, 30:3–71.

Saracevic Tefko. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. Journal of the American Society for Information Science and Technology, 58(13):2126–2144.

Scarton Carolina, Paetzold Gustavo, and Specia Lucia. 2018a. Text simplification from professionally produced corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Scarton Carolina, Paetzold Gustavo, and Specia Lucia. 2018b. Text simplification from professionally produced corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Schamber Linda, Eisenberg Michael B, and Nilan Michael S. 1990. A re-examination of relevance: toward a dynamic, situational definition. Information Processing and Management, 26(6):755–776.

See Abigail, Liu Peter J, and Manning Christopher D. 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.

Shah Deven Santosh, Schwartz H Andrew, and Hovy Dirk. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264, Online. Association for Computational Linguistics.

Jones Karen Sparck. 1998. Automatic summarising: factors and directions. arXiv e-prints, pages cmp–lg.

Synofzik Matthis, Vosgerau Gottfried, and Newen Albert. 2008. Beyond the comparator model: a multifactorial two-step account of agency. Consciousness and cognition, 17(1):219–239. [PubMed: 17482480]

Takayoshi Pamela and Selfe Cynthia L. 2007. Multimodal composition: Resources for teachers.

Tripodi Francesca. 2018. Searching for alternative facts. Data & Society.

Walker Marilyn A, Whittaker Stephen J, Stent Amanda, Maloor Preetam, Moore Johanna, Johnston Michael, and Vasireddy Gunaranjan. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. Cognitive Science, 28(5):811–840.

Walters Shannon. 2010. Toward an accessible pedagogy: Dis/ability, multimodality, and universal design in the technical communication classroom. Technical Communication Quarterly, 19(4):427–454.

Wang William Yang, Mehdad Yashar, Radev Dragomir, and Stent Amanda. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 58–68.

Wen Tsung-Hsien, Gašic Milica, Mrkšic Nikola, Su Pei-Hao, Vandyke David, and Young Steve. Semantically conditioned lstm-based natural language generation for spoken dialogue systems.

White John S., O'Connell Theresa A., and O'Mara Francis E.. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, Maryland, USA.

Witteveen Sam, Red Dragon AI, and Andrews Martin. 2019. Paraphrasing with large language models. EMNLP-IJCNLP 2019, page 215.

Yang Yi, Yih Wen-tau, and Meek Christopher. 2015. Wikiqa: A challenge dataset for open-domain question answering. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 2013–2018.

Zhang Saizheng, Dinan Emily, Urbanek Jack, Szlam Arthur, Kiela Douwe, and Weston Jason. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zhong Peixiang, Zhang Chen, Wang Hao, Liu Yong, and Miao Chunyan. 2020. Towards persona-based empathetic conversational models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6556–6566, Online. Association for Computational Linguistics.

Zhu Junnan, Li Haoran, Liu Tianshang, Zhou Yu, Zhang Jiajun, and Zong Chengqing. 2018. Msmo: Multimodal summarization with multimodal output. In Proceedings of the 2018 conference on empirical methods in natural language processing, pages 4154–4164.

Zyskind Guy, Nathan Oz, et al. 2015. Decentralizing privacy: Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops, pages 180–184. IEEE.