Towards User-Centric Text-to-Text Generation: A Survey

Diyi Yang[1] and Lucie Flek[2(✉)]

[1] School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
diyi.yang@cc.gatech.edu
[2] Conversational AI and Social Analytics (CAISA) Lab, Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany
lucie.flek@uni-marburg.de

Abstract. Natural Language Generation (NLG) has received much attention with rapidly developing models and ever-more available data. As a result, a growing amount of work attempts to personalize these systems for better human interaction experience. Still, diverse sets of research across multiple dimensions and numerous levels of depth exist and are scattered across various communities. In this work, we survey the ongoing research efforts and introduce a categorization of these under the umbrella user-centric natural language generation. We further discuss some of the challenges and opportunities in NLG personalization.

Keywords: User modeling · Personalization · NLG

1 Motivation

With an increasing output quality of text-to-text NLG models, the attention of the field is turning towards the ultimate goal, to enable human-like natural language interactions. Even outside of the dialog-system area, the generated language is produced to fulfill specific communication goals [113], hence should be tailored to the specific audience [43,100]. Human speakers naturally use a conceptual model of the recipient in order to achieve their communication goal more efficiently, for example adjust the style or level of complexity [60,101,126,150]. It is therefore reasonable to assume that such user models improve the quality of NLG systems through better adaptivity and robustness [36,82], and to personalize the system outcomes based on the available relevant information about the user. Research in this area is driven by insights from numerous disciplines, from psychology across linguistics to human-computer interaction, while the industry focus on customer-driven solutions powers the personalization of conversational assistants [8,13,14,22]. As a result, research contributions are scattered across diverse venues. Our aim is to help to limit duplicate research activities, and to organize user-centric efforts within the NLG community. The possibilities of personalizing generated text towards the user range across multiple dimensions and numerous

K. Ekštein et al. (Eds.): TSD 2021, LNAI 12848, pp. 3–22, 2021.
https://doi.org/10.1007/978-3-030-83527-9_1

levels of depth, from factual knowledge over preferences and opinions to stylistic discourse adjustments. We use for all these user adjustment variations an umbrella term *user-centric natural language generation*. We provide a comprehensive overview of recent approaches and propose a categorization of ongoing research directions.

2 Related Surveys

Related to our work, [118] conduct a survey of datasets for dialogue systems, yet noting that "personalization of dialogue systems as an important task, which so far has not received much attention". [32] surveys user profiling datasets, however, without an NLG focus. Given various input types in NLG (e.g., tables [99], RDF triple [44], meaning representation [31]), we narrow our focus to user-centric text-to-text generation when referring to user-centric NLG in this work.

3 User-Centric NLG

Generally, NLG[1] is a process that produces textual content (a sequence of consecutive words) based on a chosen structured or unstructured input. In the ideal case, such textual content shall be syntactically and semantically plausible, resembling human-written text [45,46]. NLG encompasses a wide range of application tasks [43], such as neural machine translation, text summarization, text simplification, paraphrasing with style transfer, human-machine dialog systems, video captioning, narrative generation, or creative writing [43].

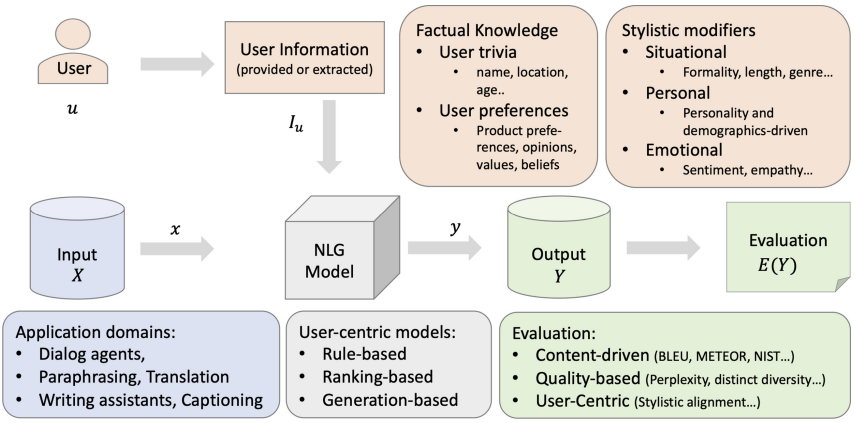

Fig. 1. User-centric natural language generation

3.1 When Is NLG User-Centric?

Given a text generation problem transforming an input x to an output y, we refer to it as **user-centric natural language generation** system when the

[1] In this work, NLG mainly refers to text-to-text generation.

output y of the NLG model is conditioned by information I_u available about the user u. In other words, the information I_u is leveraged to alter the projection of an input x to the output space. As illustrated in Fig. 1, this available user information can be of various kinds depending on specific application domains, which we categorize as follows.

3.2 User and Application Domain

In this paper we interpret the **user** in the term "user-centric" as the recipient of the generated text. Note that the previous work on personalized NLG, which we review here, sometimes takes an author-centric rather than recipient-centric view, for example dialog system works often refer to personalization as modeling of the chatbot persona [63]. The specific role of a user is dependent on a particular **application domain**, which also typically characterizes the type of the input. Below are some common application domains and the user and input examples.

- Conversational agents [69]: User is the human participant of the conversation. Input are typically the preceding utterances.
- Personalized machine translation [92,94,107,130]: User is the requester of the translation. Input is the text to be translated. Prior work mainly studied how particular personal trait of the author such as gender gets manifested in the original texts and in translations. [94] introduced a personalized machine translation system where users' preferred translation is predicted first based on similar other users.
- Writing Assistants: User is the final editor of the generated text, typically also the author of the input. Most current automated response generations such as Smart Reply in emails [65] are conducted in a generic way, not an user-specific manner.
- Personalized text simplification [9,72,89]: User is the reader, input is the text to be simplified.

Depending on whether a user is the recipient or the actor of a text, user-centric systems adapt themselves accordingly in terms of how to incorporate personalized or user-specific information into the modeling process.

Diverse Understanding of Personalization. As shown in Table 1, the interpretation of what personalized NLG means varies largely. Many systems optimize for speaker persona consistency or traits, while others operate with recipient's preferences. However, only a few studies considered recipients in their models [28]. We therefore argue, that in order to "solve" user-centric NLG, we must state more explicitly who our users are, what user needs we assume from them, and more importantly, how these user needs are reflected in our system design.

3.3 User Information

As shown in Fig. 1, we categorize user information into the following categories: (1) factual knowledge, which includes (1a) user trivia and (1b) preferences, and

(2) stylistic modifiers, which encompass (2a) situational, (2b) personal, and (2c) emotional choices.

(1) Factual Knowledge. Incorporating factual knowledge specific to a given user is essential in increasing user engagement. Information concerning **user trivia (1a)** can include personal data such as user's name or location, or user attributes such as occupation. For instance, [141] include user facts such as "*i have four kids*", although factual knowledge is not introduced in a structured way. [19] utilized product user categories to generate personalized product descriptions. [91] uses reinforcement learning to reward chatbot persona consistency using fact lists with information like "*my dad is a priest*". User facts

Table 1. Overview table of example previous user-centric NLG works that fall into each user information type. Note that the majority of works focuses on modeling the speaker persona rather than personalizing towards a representation of a recipient.

Research work	NLG task	Input X	User u	User info. I_u	I_u example	NLG model
[141]	ConvAgent (chitchat)	Utterances (PERSONA-CHAT)	Person being talked to	Factual - trivia, - preferences	Family, job, hobbies	Memory network
[19]	Product description	Product title (E-commerce)	Target customer	Factual - preferences	Category, aspect focus	Transformer
[101]	Device description	Device patents	Knowledge seeker	Factual - trivia	Background knowledge	Rule-based
[114]	ConvAgent (chitchat)	Argumentative interaction (Kialo Dataset)	Speaker persona	Factual -preferences	Stances, beliefs	Seq2seq
[63]	ConvAgent (goal-oriented)	bAbI dialog	Speaker persona	Style - personal	Age, gender	Memory network
[84]	ConvAgent (goal-oriented)	bAbI dialog	Speaker persona	Factual - preferences	Pref. over KB (embeddings)	Embedding Memory network
[75]	Chitchat	OpenSubtitles	Speaker persona	Style - situational	Specificity	Seq2seq, RL Data distillation
[86]	ConvAgent (chitchat)	PERSONAGE	Speaker persona	Style - personal	Big 5 Personality traits	Rule-based
[50]	ConvAgent (chitchat)	PERSONAGE	Speaker persona	Style - personal	Big 5 personality traits	Seq2seq
[97]	ConvAgent (chitchat)	Restaurant utterances	Speaker persona	Style - personal	Personality traits	Seq2seq
[74]	ConvAgent (chitchat)	Twitter, TV	Speaker, Recipient	Factual, Style (all)	Embeddings	Speaker model
[28]	ConvAgent (healthcare)	PTDS healthcare	Speaker, Recipient	Style - situational	Verbal, Non-verbal	Rule-based
[41]	ConvAgent (chitchat)	Share emotions	Speaker, Recipient	Style - emotional	Empathy	Rule-based
[59]	ConvAgent (chitchat)	Prior utterances	Speaker persona	Style - personal	Personality	N-gram LM
[148]	ConvAgent (chitchat)	Weibo	Speaker persona	Style - emotional	Emotion	Seq2seq
[42]	Reader-aware summarization	Weibo	Speaker, Recipient	Factual - preferences	Opinion	Seq2seq

for personalization include also expertise level in tutoring systems [53,101]. In addition to user trivia, including **user preferences (1b)** ("*i hate Mexican food*" or "*i like to ski*" [141]), such as opinions, values, and beliefs [114] has been of importance for dialog systems, as it leads to producing more coherent and interesting conversations.

(2) Stylistic Modifiers. Stylistic variation can be characterized as a variation in phonological, lexical, syntactic, or discourse realisation of a particular semantic content, due to user's characteristics [15,45]. To date, most of the style adaptation work in the NLG area focused on the **situational stylistic modifiers (2a)**, perceiving language use as a function of intended audience, genre, or situation. For example, professional/colloquial [35], personal/impersonal, formal/informal [18,96,102,103,110,136,143] or polite/impolite [25,38,85,95,116]. Recently, unsupervised style transfer has gained popularity [70].

Comparably less research has been conducted in the emotional and personal modifiers, such as empathy or demographics. **Personal stylistic modifiers (2b)** in our scheme include user attributes, i.e. both conscious and unconscious traits intrinsic to the text author's individual identity [59]. A common property of these traits is that while their description is typically clear, such as *teenager*, *Scottish*, or *extrovert*, their surface realization is less well-defined [7]. Note that this is different from employing these attributes as user trivia in a factual way. The two main subgroups of personal modifiers are **sociodemographic traits** and **personality**. NLG words explore mostly gendered paraphrasing and gender-conditioned generation [104,105,112,127]. Personality has been employed mostly in the dialog area, mainly on the agent side [50,95,97]. In an early work on personality-driven NLG, the system of [87] estimates generation parameters for stylistic features based on the input of big five personality traits [24]. For example, an introverted style may include more hedging and be less verbose. While the big five model is the most widely accepted in terms of validity, its assessments are challenging to obtain [123]. Some works thus resort to other personality labels [41,132], or combinations of sociodemographic traits and personal interests [146]. Modeling personality of the recipient of the generated text is rare in recent NLG systems, although it has been shown to affect e.g. argument persuasiveness [33,83] and capability of learning from a dialog [26]. For example [53] proposed to use a multi-dimensional user model including hearer's emotional state and interest in the discussion, [26] represented users' stylistic preference for verboseness and their discourse understanding ability, and [11] inferred user's psychological states from their actions to update the model of a user's beliefs and goals. [55] uses LIWC keywords to infer both instructor's and recipient's personality traits to study dialog adaptation.

Emotional stylistic modifiers (2c) encompass the broad range of research in the area of affective NLG [26]. In the early works, manually prepared rules are applied to deliberately select the desired emotional responses [124], and pattern-based models are used to generate text to express emotions [66]. There is a broad range of features beyond affective adjectives that can have emotional impact, such as an increased use of redundancy, first-person pronouns, and adverbs [27]. [47]

introduce neural language models which allows to customize the degree of emotional content in generated sentences through an additional design parameter (happy, angry, sad, anxious, neutral). They note that it is difficult to produce emotions in a natural and coherent way due to the required balance of grammatically and emotional expressiveness. [4] show three novel ways to incorporate emotional aspects into encoder-decoder neural conversation models: word embeddings augmented with affective dictionaries, affect-based loss functions, and affectively diverse beam search for decoding. In their work on emotional chatting machines, [148] demonstrates that simply embedding emotion information in existing neural models cannot produce desirable emotional responses but just ambiguous general expressions with common words. They proposes a mechanism, which, after embedding emotion categories, captures the change of implicit internal emotion states, and boosts the probability of explicit emotion expressions with an external vocabulary. [125] observe, in line with [27], that one doesn't need to explicitly use strong emotional words to express emotional states, but one can implicitly combine neutral words in distinct ways to increase the intensity of the emotional experiences. They develop two NLG models for emotionally-charged text, explicit and implicit. The ability to produce language controlled for emotions is closely tied to the goal of building empathetic social chatbots [28,40,41,111,121]. To date, these mainly leverage emotional embeddings similar to those described above to generate responses expected by the learned dialog policies. [78] point out the responses themselves don't need to be emotional, but mainly understanding, and propose a model based on empathetic listeners.

Implicit User Modeling. With the rise of deep learning models and the accompanying learned latent representations, boundaries between the user information categories sometimes get blurred, as the knowledge extracted about the user often isn't explicitly interpreted. This line of work uses high-dimensional vectors to refer to different aspects associated with users, implicitly grouping users with similar features (whether factual or stylistic) into similar areas of the vector space. Neural user embeddings in the context of dialog modeling have been introduced by [74], which capture latent speaker persona vectors based on speaker ID. This approach has been further probed and enhances by many others [63], e.g. by pretraining speaker embeddings on larger datasets [69,137,146,147], incorporating user traits into the decoding stage [145], or via mutual attention [88].

4 Data for User-Centric NLG

We identify five main types of datasets that can be leveraged for user-centric NLG, and provide their overview in Table 2. These types include: (1) Attribute-annotated datasets for user profiling, such as in [109], (2) style transfer and attribute transfer paraphrasing datasets such as [110], (3) attribute-annotated machine translation datasets such as [130], (4) persona-annotated dialog datasets such as [141], and (5) large conversational or other human-generated datasets with speaker ID, which allow for unsupervised speaker representation training.

Table 2. Available datasets usable for user-centric NLG

Task	Data and size	User info
[6] Dialog modeling	Movie dialogs, 132K conv.	Speaker ID
[69] Dialog modeling	Movie dialogs, 4.4K conv.	Speaker ID
[3] Character modeling	Movie subtitles, 5.5M turn pairs	Speaker ID
[141] Persona modeling	Chit-chat, 1K pers	Persona traits
[2] User modeling	Reddit, 133M conv 2.1B posts	User data
[74] Persona modeling	Twitter, 74K users (24M conv.)	User data
[146] Persona modeling	Weibo, 8.47M users, 21M conv.	Gender, age, loc.
[134] Attribute transfer	Reddit, Facebook, ≥100K posts	Political Slant
[112] Attribute transfer	Twitter, Yelp, ≥1M users	Gender
[105] Attribute transfer	Words, phrases	Gender
[110] Style transfer	Yahoo Answers, 110K pairs	Formality
[85] Style transfer	Enron e-mails, 1.39M texts	Politeness
[140] Style transfer	Twitter, 14K Tweets	Offensiveness
[10] Attribute transfer	Product QA ≥9K quest	Subjectivity
[76] Attribute transfer	≥1M reviews	Sentiment
[92] Machine translation	TED talks, 2.3K Talks (271K sent.)	Speaker
[107] Machine translation	EuroParl, ≥100K sent. pairs (de, fr)	Gender
[130] Machine translation	EuroParl, ≥100K pairs (20 lang.)	Gender, age

In addition, as [12] point out, the challenge in the big data era is not to find human generated dialogues, but to employ them appropriately for social dialogue generation. Any existing social media dialogues can be combined with a suite of tools for sentiment analysis, topic identification, summarization, paraphrase, and rephrasing, to bootstrap a socially-apt NLG system.

5 User-Centric Generation Models

Already [150] discuss how natural language systems consult user models in order to improve their understanding of users' requirement and to generate appropriate and relevant responses. Generally, current user-centric generation models can be divided into rule-based, ranking-based and generation-based models.

Rule-based user models often utilize a pre-defined mapping between user types and topics [34], or hand-crafted user and context features [1]. The recent Alexa Prize social-bots also utilized a pre-defined mapping between personality types and topics [34], or hand-crafted user and context features [1].

Ranking-based models [2,90,141] focus on the task of response selection from a pool of candidates. Such response selection relies heavily on learning the matching between the given user post and any response from the pool, such

as the deep structured similarity models [56] or the deep attention matching network [149]. [80] proposed to address the personalized response ranking task by incorporating user profiles into the conversation model. Generation-based models attempt to generate response directly from any given input questions. Most widely used models are built upon sequence-to-sequence models, and the recent transformer-based language models pretrained with large corpora [144].

With the development of large scale social media data [69,117,119,128,145], several personalized response generation models have been proposed. [21] introduced a neural model to learn a dynamically updated speaker embedding in a conversational context. They initialized speaker embedding in an unsupervised way by using context-sensitive language generation as an objective, and fine-tuned it at each turn in a dialog to capture changes over time and improve the speaker representation with added data. [74] introduced the Speaker Model that encoded user-id information into an additional vector and fed it into the decoder to capture the identity of the speakers. In addition to using user id to capture personal information, [141] proposed a profile memory network for encoding persona sentences. Recently, there are a few works using meta-learning and reinforcement learning to enhance mutual persona perception [68,79,88]. Generative models can produce novel responses, but they might suffer from grammar errors, repetitive, hallucination, and even uncontrollable outputs, all of which might degrade the performance of user-centric generation. For instance, under personalized dialog settings, [141] claimed that ranking-based models performed better than generative models, suggesting that building user-centric generative models is more challenging.

Hybrid models attempt to combine the strengths of the generative and rank paradigms [138] in a two-stage fashion, i.e., retrieving similar or template responses first, and then using these to help generate new responses. Hybrid models shed light on how to build user-centric NLG models as the first stage can be used to retrieve relevant knowledge/responses and the second stage can fine-tune the retrieved ones to be user-specific.

6 Evaluations

Current automatic evaluation metrics for response generation can be broadly categorized into three classes: content-driven, quality-based and user-centric. **Content** relatedness measures capture the distance of the generated response from its corresponding ground-truth, with representative metrics such as BLEU [98], NIST [30], and METEOR [71]. Speaker sensitive responses evaluation model [5] enhances the relatedness score with a context-response classifier. From a **quality** perspective, the fluency and diversity matter, assessed via perplexity [20] and distinct diversity [73]. From a **user-centric** perspective, we need to evaluate the style matching or fact adherence that compare the generated responses' language to the user's own language. Existing example metrics include the stylistic alignment [93,129] at the surface, lexical and syntactic level, model-driven metrics such as Hits@1/N, calculating how accurate the generated response can be

automatically classified to its corresponding user or user group [29,93], and the average negative log-likelihood of generated text to user-specific language model, e.g. for poet's lyrics [131].

Evaluation towards open-ended conversations [64,106] also use Grice's Maxims of Conversation [49], i.e., evaluating whether the generated content violates *Quantity* that gives more or less information than requires, *Quality* that shares false information or things we do not have evidence, *Relation* that stays on the relevant topic, and *Manner* that requires communicating clearly without much disfluency. [67] further introduced a new diagnostic measure called relative utterance quantity (RUQ) to see if the model favors a generic response (e.g., '*I don't know*'). over the reference it was trained on.

Despite various measures in automatically assessing the quality of responses generated, human evaluation still plays a key role in assessing user-centric NLG systems, as the correlation between automated and human quality judgments is very limited [81,93]. Automatic metrics for evaluating user-centric NLG systems could then come in the form of an evaluation model learned from human data, e.g. collected from surveys, in order to provide human-like scores to proposed responses like BLEURT [115]. Recently, [54] argued that although human assessment remains the most trusted form of evaluation, the NLG community takes highly diverse approaches and different quality criteria, making it difficult to compare results and draw conclusions, with adverse implications for meta-evaluation and reproducibility. Their analyses on top of 165 NLG papers call for standard methods and terminology for NLG evaluation.

Human judgement for user-centric NLG requires significant efforts. User information such as styles, opinions or personalized knowledge is often scattered throughout the entire participation history in various formats such as posts, comments, likes or log-ins. It is impossible for annotators to go through these hundreds of activity records to infer whether the generated response fits the user well; furthermore, personalization is hardly reflected in a single message, but mostly inferred from a large collection of users' activities. [123]. Moreover, users' preferences and interests change over time either slowly or rapidly [48,77], making it even harder to third-parties to judge and evaluate. As a result, direct and self-evaluation from users of the user-centric NLG systems deserves more attention.

7 Challenges and Opportunities

User-Centric Data Collection and Evaluation. Collecting large-scale personalized conversations or data for NLG systems is challenging, expensive and cumbersome. First, most datasets suffer from pre-defined or collected user profiles expressed in a limited number of statements. Second, crowdsourcing personalized datasets is likely to result in very artificial content, as the workers need to intentionally inject the received personalization instructions into every utterance, which does not align well with human daily conversations. Correspondingly, state-of-the-art models tend to perform the attribute transfer merely at the lexical level (e.g. inserting negative words for *rude* or "please" for *polite*), while the

subtle understanding and modification of higher-level compositionality is still a challenge [39,62,148]. Even more problematic assumption of most user-centric generation systems is that users exhibit their traits, moods and beliefs uniformly in a conversation. However, humans do not always express personalized information everywhere, thus real world data is persona-sparse [147]. This calls for a nuanced modeling of when, where and to what extent personalization needs to be included for NLG systems [17,37,122].

Personalized Pretraining and Safeguards. Getting data is a key challenge when it comes to personalized pre-training [147], which requires extensive data even for each single user. The proliferation of personalization also brings in trust and privacy issues [23,120]. How does user-centric generation relate to ethics and privacy as the personalization always involve using user specific data [51]? One key issue associated with personalized pretraining is that the extensive personal data needed by pretrained language models might include all sorts of dimensions about users, including sensitive and private information which should not be used by user-centric NLG systems [52,108]. For instance, [16] demonstrated that an adversary can perform an extraction attack to recover individual training examples from pretrained language models. These extracted examples include names, phone numbers, email addresses, and even deleted content. Such privacy concerns might become more salient and severe when it comes to user-centric pretraining, as models can easily remember details and leak training data for potential malicious attacks.

Biases and Generalization. The creation of corpora for user-centric NLG might suffer from *self-selection bias* as people who decides to use certain platforms like Twitter or Reddit might be very different. The *reporting bias* further adds complexity to this space as people do not necessarily talk about things in the world in proportion to their persona or personality. Thus, NLG systems built upon available data might be skewed towards certain population (e.g., educational background, access to Internet, specific language uses). The *crowdsourcing bias* [57], i.e., the potential inherent bias of crowd workers who contribute to the tasks might introduce biased ground-truth data.

Gaps Between Users and Systems. We argue that the evaluation process should look into what dimension users expect to see and identify what users want from these generated texts. For example [135] points out the expectations from human and artificial participants of the conversation are not the same, and shall be modeled differently. We need metrics to capture any failures, and mechanisms to explain the decision-making process behind these user-centric NLG models, since the data-driven systems tend to imitate utterances from their training data [61,133,139]. This process is not directly controllable, which may lead to offensive responses [58]. Another challenge is how to disentangle personalization from the generic representation [39], such as using domain adaptation techniques to transfer generic models to specific user groups [142].

8 Conclusion

This work presents a comprehensive overview of recent user-centric text generation across diverse sets of research capturing multiple dimensions of personalizing systems. We categorize these previous research directions, and present the representative tasks and evaluation approaches, as well as challenges and opportunities to facilitate future work on user-centric NLG.

References

1. Ahmadvand, A., et al.: Emory irisbot: an open-domain conversational bot for personalized information access. In: Alexa Prize Proceedings (2018)
2. Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y., Strope, B., Kurzweil, R.: Conversational contextual cues: the case of personalization and history for response ranking. arXiv preprint arXiv:1606.00372 (2016)
3. Ameixa, D., Coheur, L., Fialho, P., Quaresma, P.: Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In: Bickmore, T., Marsella, S., Sidner, C. (eds.) IVA 2014. LNCS (LNAI), vol. 8637, pp. 13–21. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09767-1_2
4. Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L.: Affective neural response generation. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 154–166. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_12
5. Bak, J., Oh, A.: Speaker sensitive response evaluation model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6376–6385. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.568. https://www.aclweb.org/anthology/2020.acl-main.568
6. Banchs, R.E.: Movie-DiC: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 203–207 (2012)
7. Belz, A.: ITRI-03-21 and now with feeling: developments in emotional language generation (2003)
8. Biller, M., Konya-Baumbach, E., Kuester, S., von Janda, S.: Chatbot anthropomorphism: a way to trigger perceptions of social presence? In: Blanchard, S. (ed.) 2020 AMA Summer Academic Conference: Bridging Gaps: Marketing in an Age of Disruption, vol. 31, pp. 34–37. American Marketing Association, Chicago (2020). https://madoc.bib.uni-mannheim.de/56482/
9. Bingel, J., Paetzold, G., Søgaard, A.: Lexi: a tool for adaptive, personalized text simplification. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 245–258 (2018)
10. Bjerva, J., Bhutani, N., Golshan, B., Tan, W.C., Augenstein, I.: SubjQA: a dataset for subjectivity and review comprehension. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5480–5494 (2020)
11. Bonarini, A.: Modeling issues in multimedia car-driver interaction. In: Proceedings of the 1991 International Conference on Intelligent Multimedia Interfaces, pp. 353–371 (1991)

12. Bowden, K.K., Oraby, S., Misra, A., Wu, J., Lukin, S., Walker, M.: Data-driven dialogue systems for social agents. In: Eskenazi, M., Devillers, L., Mariani, J. (eds.) Advanced Social Interaction with Agents. LNEE, vol. 510, pp. 53–56. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-92108-2_6
13. Bowden, K.K., et al.: Entertaining and opinionated but too controlling: a large-scale user study of an open domain Alexa prize system. In: Proceedings of the 1st International Conference on Conversational User Interfaces, pp. 1–10 (2019)
14. Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., Alt, F.: At your service: designing voice assistant personalities to improve automotive user interfaces. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2019)
15. Brooke, J., Flekova, L., Koppel, M., Solorio, T.: Proceedings of the Second Workshop on Stylistic Variation (2018)
16. Carlini, N., et al.: Extracting training data from large language models. arXiv preprint arXiv:2012.07805 (2020)
17. Chaves, A.P., Gerosa, M.A.: How should my chatbot interact? A survey on human-chatbot interaction design. arXiv preprint arXiv:1904.02743 (2019)
18. Chawla, K., Srinivasan, B.V., Chhaya, N.: Generating formality-tuned summaries using input-dependent rewards. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 833–842. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/K19-1078. https://www.aclweb.org/anthology/K19-1078
19. Chen, Q., Lin, J., Zhang, Y., Yang, H., Zhou, J., Tang, J.: Towards knowledge-based personalized product description generation in e-commerce. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3040–3050 (2019)
20. Chen, S.F., Beeferman, D., Rosenfeld, R.: Evaluation metrics for language models (1998)
21. Cheng, H., Fang, H., Ostendorf, M.: A dynamic speaker model for conversational interactions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2772–2785 (2019)
22. Churamani, N., et al.: The impact of personalisation on human-robot interaction in learning scenarios. In: Proceedings of the 5th International Conference on Human Agent Interaction, pp. 171–180 (2017)
23. Coavoux, M., Narayan, S., Cohen, S.B.: Privacy-preserving neural representations of text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1–10. Association for Computational Linguistics, Brussels, October–November 2018. https://doi.org/10.18653/v1/D18-1001. https://www.aclweb.org/anthology/D18-1001
24. Costa, P.T., Jr., McCrae, R.R.: Personality disorders and the five-factor model of personality. J. Pers. Disord. **4**(4), 362–371 (1990)
25. Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., Potts, C.: A computational approach to politeness with application to social factors. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 250–259. Association for Computational Linguistics, Sofia, August 2013. https://www.aclweb.org/anthology/P13-1025
26. de Rosis, F., Grasso, F.: Affective natural language generation. In: Paiva, A. (ed.) IWAI 1999. LNCS (LNAI), vol. 1814, pp. 204–218. Springer, Heidelberg (2000). https://doi.org/10.1007/10720296_15

27. De Rosis, F., Grasso, F., Castelfranchi, C., Poggi, I.: Modelling conflict-resolution dialogues. In: Müller, H.J., Dieng, R. (eds.) Computational Conflicts, pp. 41–62. Springer, Heidelberg (2000). https://doi.org/10.1007/978-3-642-56980-7_3
28. DeVault, D., et al.: SimSensei Kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, pp. 1061–1068 (2014)
29. Dinan, E., et al.: The second conversational intelligence challenge (ConvAI2). arXiv preprint arXiv:1902.00098 (2019)
30. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145 (2002)
31. Dušek, O., Howcroft, D.M., Rieser, V.: Semantic noise matters for neural natural language generation. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 421–426 (2019)
32. Eke, C.I., Norman, A.A., Shuib, L., Nweke, H.F.: A survey of user profiling: state-of-the-art, challenges, and solutions. IEEE Access **7**, 144907–144924 (2019)
33. El Baff, R., Al Khatib, K., Stein, B., Wachsmuth, H.: Persuasiveness of news editorials depending on ideology and personality. In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, pp. 29–40 (2020)
34. Fang, H., et al.: Sounding board-university of Washington's Alexa prize submission. In: Alexa Prize Proceedings (2017)
35. Ficler, J., Goldberg, Y.: Controlling linguistic style aspects in neural language generation. In: Proceedings of the Workshop on Stylistic Variation, pp. 94–104. Association for Computational Linguistics, Copenhagen, September 2017. https://doi.org/10.18653/v1/W17-4912. https://www.aclweb.org/anthology/W17-4912
36. Finin, T.W.: GUMS-a general user modeling shell. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems. SYMBOLIC, pp. 411–430. Springer, Heidelberg (1989). https://doi.org/10.1007/978-3-642-83230-7_15
37. Flek, L.: Returning the N to NLP: towards contextually personalized classification models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7828–7838 (2020)
38. Fu, L., Fussell, S., Danescu-Niculescu-Mizil, C.: Facilitating the communication of politeness through fine-grained paraphrasing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5127–5140. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.416. https://www.aclweb.org/anthology/2020.emnlp-main.416
39. Fu, Y., Zhou, H., Chen, J., Li, L.: Rethinking text attribute transfer: a lexical analysis. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 24–33 (2019)
40. Fung, P., Bertero, D., Xu, P., Park, J.H., Wu, C.S., Madotto, A.: Empathetic dialog systems. In: LREC 2018 (2018)
41. Fung, P., et al.: Zara the supergirl: an empathetic personality recognition system. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 87–91 (2016)
42. Gao, S., et al.: Abstractive text summarization by incorporating reader comments. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6399–6406 (2019)
43. Garbacea, C., Mei, Q.: Neural language generation: formulation, methods, and evaluation. arXiv preprint arXiv:2007.15780 (2020)

44. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: The WebNLG challenge: generating text from RDF data. In: Proceedings of the 10th International Conference on Natural Language Generation, pp. 124–133 (2017)
45. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018)
46. Gehrmann, S., et al.: The gem benchmark: natural language generation, its evaluation and metrics. arXiv preprint arXiv:2102.01672 (2021)
47. Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S.: 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), pp. 634–642 (2017)
48. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science **333**(6051), 1878–1881 (2011)
49. Grice, H.P.: Logic and conversation. In: Speech Acts, pp. 41–58. Brill (1975)
50. Harrison, V., Reed, L., Oraby, S., Walker, M.: Maximizing stylistic control and semantic accuracy in NLG: personality variation and discourse contrast. In: Proceedings of the 1st Workshop on Discourse Structure in Neural NLG, pp. 1–12 (2019)
51. Henderson, P., et al.: Ethical challenges in data-driven dialogue systems. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 123–129 (2018)
52. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep models under the GAN: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 603–618 (2017)
53. Hovy, E.: Generating natural language under pragmatic constraints. J. Pragmat. **11**(6), 689–719 (1987)
54. Howcroft, D.M., et al.: Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In: Proceedings of the 13th International Conference on Natural Language Generation, pp. 169–182 (2020)
55. Hu, Z., Tree, J.E.F., Walker, M.: Modeling linguistic and personality adaptation for natural language generation. In: Proceedings of the 19th annual SIGdial Meeting on Discourse and Dialogue, pp. 20–31 (2018)
56. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2333–2338 (2013)
57. Hube, C., Fetahu, B., Gadiraju, U.: Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
58. Hunt, E.: Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. The Guardian, 24 March 2016. http://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter
59. Isard, A., Brockmann, C., Oberlander, J.: Individuality and alignment in generated dialogues. In: Proceedings of the Fourth International Natural Language Generation Conference, pp. 25–32 (2006)
60. Jameson, A.: But what will the listener think? Belief ascription and image maintenance in dialog. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems. SYMBOLIC, pp. 255–312. Springer, Heidelberg (1989). https://doi.org/10.1007/978-3-642-83230-7_10

61. Ji, Z., Lu, Z., Li, H.: An information retrieval approach to short text conversation. arXiv:1408.6988 [cs], August 2014. http://arxiv.org/abs/1408.6988
62. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: a survey (2020)
63. Joshi, C.K., Mi, F., Faltings, B.: Personalization in goal-oriented dialog. arXiv preprint arXiv:1706.07503 (2017)
64. Jwalapuram, P.: Evaluating dialogs based on Grice's maxims. In: Proceedings of the Student Research Workshop Associated with RANLP, pp. 17–24 (2017)
65. Kannan, A., et al.: Smart reply: automated response suggestion for email. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 955–964 (2016)
66. Keshtkar, F., Inkpen, D.: A pattern-based model for generating text to express emotion. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6975, pp. 11–21. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24571-8_2
67. Khayrallah, H., Sedoc, J.: Measuring the 'i don't know' problem through the lens of Gricean quantity. arXiv preprint arXiv:2010.12786 (2020)
68. Kim, H., Kim, B., Kim, G.: Will i sound like me? Improving persona consistency in dialogues through pragmatic self-consciousness (2020)
69. Kottur, S., Wang, X., Carvalho, V.: Exploring personalized neural conversational models. In: IJCAI, pp. 3728–3734 (2017)
70. Krishna, K., Wieting, J., Iyyer, M.: Reformulating unsupervised style transfer as paraphrase generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 737–762 (2020)
71. Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231 (2007)
72. Lee, J.S., Yeung, C.Y.: Personalizing lexical simplification. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 224–232 (2018)
73. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
74. Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B.: A persona-based neural conversation model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 994–1003 (2016)
75. Li, J., Monroe, W., Jurafsky, D.: Data distillation for controlling specificity in dialogue generation. arXiv preprint arXiv:1702.06703 (2017)
76. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1865–1874 (2018)
77. Li, L., Zheng, L., Yang, F., Li, T.: Modeling and broadening temporal user interest in personalized news recommendation. Expert Syst. Appl. **41**(7), 3168–3177 (2014)
78. Lin, Z., Madotto, A., Shin, J., Xu, P., Fung, P.: 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 (EMNLP-IJCNLP), pp. 121–132 (2019)
79. Lin, Z., Madotto, A., Wu, C.S., Fung, P.: Personalizing dialogue agents via meta-learning (2019)

80. Liu, B., et al.: Content-oriented user modeling for personalized response ranking in chatbots. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 122–133 (2017)
81. Liu, C.W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016)
82. Lucas, J., Fernández, F., Salazar, J., Ferreiros, J., San Segundo, R.: Managing speaker identity and user profiles in a spoken dialogue system. Procesamiento del lenguaje natural (43), 77–84 (2009)
83. Lukin, S., Anand, P., Walker, M., Whittaker, S.: Argument strength is in the eye of the beholder: audience effects in persuasion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 742–753 (2017)
84. Luo, L., Huang, W., Zeng, Q., Nie, Z., Sun, X.: Learning personalized end-to-end goal-oriented dialog. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6794–6801 (2019)
85. Madaan, A., et al.: Politeness transfer: a tag and generate approach. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1869–1881 (2020)
86. Mairesse, F., Walker, M.: PERSONAGE: personality generation for dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 496–503 (2007)
87. Mairesse, F., Walker, M.: Trainable generation of big-five personality styles through data-driven parameter estimation. In: Proceedings of ACL-2008: HLT, pp. 165–173 (2008)
88. Majumder, B.P., Jhamtani, H., Berg-Kirkpatrick, T., McAuley, J.: Like hiking? You probably enjoy nature: persona-grounded dialog with commonsense expansions (2020)
89. Mallinson, J., Lapata, M.: Controllable sentence simplification: employing syntactic and lexical constraints. arXiv preprint arXiv:1910.04387 (2019)
90. Mazare, P.E., Humeau, S., Raison, M., Bordes, A.: Training millions of personalized dialogue agents. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2775–2779 (2018)
91. Mesgar, M., Simpson, E., Wang, Y., Gurevych, I.: Generating persona-consistent dialogue responses using deep reinforcement learning. arXiv-2005 (2020)
92. Michel, P., Neubig, G.: Extreme adaptation for personalized neural machine translation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 312–318 (2018)
93. Mir, R., Felbo, B., Obradovich, N., Rahwan, I.: Evaluating style transfer for text. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 495–504. Association for Computational Linguistics, Minneapolis, June 2019. https://doi.org/10.18653/v1/N19-1049. https://www.aclweb.org/anthology/N19-1049
94. Mirkin, S., Meunier, J.L.: Personalized machine translation: predicting translational preferences. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2019–2025 (2015)
95. Niu, T., Bansal, M.: Polite dialogue generation without parallel data. Trans. Assoc. Comput. Linguist. **6**, 373–389 (2018). https://www.aclweb.org/anthology/Q18-1027

96. Niu, X., Martindale, M., Carpuat, M.: A study of style in machine translation: controlling the formality of machine translation output. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2814–2819. Association for Computational Linguistics, Copenhagen, September 2017. https://doi.org/10.18653/v1/D17-1299. https://www.aclweb.org/anthology/D17-1299
97. Oraby, S., Reed, L., Tandon, S., Sharath, T., Lukin, S., Walker, M.: Controlling personality-based stylistic variation with neural natural language generators. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pp. 180–190 (2018)
98. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
99. Parikh, A., et al.: ToTTo: a controlled table-to-text generation dataset. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1173–1186 (2020)
100. Paris, C.: User Modelling in Text Generation. Bloomsbury Publishing, London (2015)
101. Paris, C.L.: The use of explicit user models in a generation system for tailoring answers to the user's level of expertise. In: Kobsa, A., Wahlster, W. (eds.) User Models in Dialog Systems. SYMBOLIC, pp. 200–232. Springer, Heidelberg (1989). https://doi.org/10.1007/978-3-642-83230-7_8
102. Pavlick, E., Tetreault, J.: An empirical analysis of formality in online communication. Trans. Assoc. Comput. Linguist. **4**, 61–74 (2016). https://www.aclweb.org/anthology/Q16-1005
103. Peterson, K., Hohensee, M., Xia, F.: Email formality in the workplace: a case study on the Enron corpus. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 86–95. Association for Computational Linguistics, Portland, June 2011. https://www.aclweb.org/anthology/W11-0711
104. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. arXiv preprint arXiv:1804.09000 (2018)
105. Preotiuc-Pietro, D., Xu, W., Ungar, L.: Discovering user attribute stylistic differences via paraphrasing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
106. Qwaider, M.R., Freihat, A.A., Giunchiglia, F.: TrentoTeam at SemEval-2017 task 3: an application of Grice maxims in ranking community question answers. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 271–274 (2017)
107. Rabinovich, E., Mirkin, S., Patel, R.N., Specia, L., Wintner, S.: Personalized machine translation: preserving original author traits. arXiv preprint arXiv:1610.05461 (2016)
108. Ramaswamy, S., Thakkar, O., Mathews, R., Andrew, G., McMahan, H.B., Beaufays, F.: Training production language models without memorizing user data. arXiv preprint arXiv:2009.10031 (2020)
109. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pp. 352–365. CELCT (2013)

110. Rao, S., Tetreault, J.: Dear sir or madam, may i introduce the GYAFC dataset: corpus, benchmarks and metrics for formality style transfer. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 129–140. Association for Computational Linguistics, New Orleans, June 2018. https://doi.org/10.18653/v1/N18-1012. https://www.aclweb.org/anthology/N18-1012
111. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: a new benchmark and dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5370–5381 (2019)
112. Reddy, S., Knight, K.: Obfuscating gender in social media writing. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 17–26 (2016)
113. Reiter, E.: Natural language generation challenges for explainable AI. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pp. 3–7 (2019)
114. Scialom, T., Tekiroğlu, S.S., Staiano, J., Guerini, M.: Toward stance-based personas for opinionated dialogues. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 2625–2635 (2020)
115. Sellam, T., Das, D., Parikh, A.: BLEURT: learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892 (2020)
116. Sennrich, R., Haddow, B., Birch, A.: Controlling politeness in neural machine translation via side constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 35–40 (2016)
117. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
118. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:1512.05742 (2015)
119. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation (2015)
120. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
121. Shum, H.Y., He, X., Li, D.: From Eliza to Xiaoice: challenges and opportunities with social chatbots. Front. Inf. Technol. Electron. Eng. **19**(1), 10–26 (2018)
122. Shumanov, M., Johnson, L.: Making conversations with chatbots more personalized. Comput. Hum. Behav. **117**, 106627 (2020)
123. Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12516–12526 (2019)
124. Skowron, M.: Affect listeners: acquisition of affective states by means of conversational systems. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Development of Multimodal Interfaces: Active Listening and Synchrony. LNCS, vol. 5967, pp. 169–181. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12397-9_14

125. Song, Z., Zheng, X., Liu, L., Xu, M., Huang, X.J.: Generating responses with a specific emotion in dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3685–3695 (2019)
126. Su, P., Wang, Y.B., Yu, T., Lee, L.: A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8213–8217. IEEE (2013)
127. Subramanian, S., Lample, G., Smith, E.M., Denoyer, L., Ranzato, M., Boureau, Y.L.: Multiple-attribute text style transfer. arXiv preprint arXiv:1811.00552 (2018)
128. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in NIPS (2014)
129. Syed, B., Verma, G., Srinivasan, B.V., Natarajan, A., Varma, V.: Adapting language models for non-parallel author-stylized rewriting. In: AAAI, pp. 9008–9015 (2020)
130. Vanmassenhove, E., Hardmeier, C., Way, A.: Getting gender right in neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3003–3008 (2018)
131. Vechtomova, O., Bahuleyan, H., Ghabussi, A., John, V.: Generating lyrics with variational autoencoder and multi-modal artist embeddings. arXiv preprint arXiv:1812.08318 (2018)
132. Verhoeven, B., Daelemans, W., Plank, B.: TwiSty: a multilingual Twitter stylometry corpus for gender and personality profiling. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1632–1637. European Language Resources Association (ELRA), Portorož, May 2016. https://www.aclweb.org/anthology/L16-1258
133. Vinyals, O., Le, Q.: A neural conversational model. In: Proceedings of the 31st International Conference on Machine Learning, Lille, France, June 2015. arXiv: 1506.05869
134. Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., Tsvetkov, Y.: RtGender: a corpus for studying differential responses to gender. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
135. Völkel, S.T., et al.: Developing a personality model for speech-based conversational agents using the psycholexical approach. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2020)
136. Wang, Y., Wu, Y., Mou, L., Li, Z., Chao, W.: Harnessing pre-trained neural networks with rules for formality style transfer. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3573–3578. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/D19-1365. https://www.aclweb.org/anthology/D19-1365
137. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: TransferTransfo: a transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149 (2019)
138. Wu, Y., Wei, F., Huang, S., Wang, Y., Li, Z., Zhou, M.: Response generation by context-aware prototype editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7281–7288 (2019)

139. Yu, Z., Papangelis, A., Rudnicky, A.: TickTock: a non-goal-oriented multimodal dialog system with engagement awareness. In: Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium, Palo Alto, CA, USA, pp. 108–111 (2015). https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/viewFile/10315/10119
140. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 task 6: identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983 (2019)
141. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: 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), pp. 2204–2213. Association for Computational Linguistics, Melbourne, July 2018. https://doi.org/10.18653/v1/P18-1205. https://www.aclweb.org/anthology/P18-1205
142. Zhang, W.N., Zhu, Q., Wang, Y., Zhao, Y., Liu, T.: Neural personalized response generation as domain adaptation. World Wide Web **22**(4), 1427–1446 (2019). https://doi.org/10.1007/s11280-018-0598-6
143. Zhang, Y., Ge, T., Sun, X.: Parallel data augmentation for formality style transfer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3221–3228. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.294. https://www.aclweb.org/anthology/2020.acl-main.294
144. Zhang, Y., et al.: DialoGPT: large-scale generative pre-training for conversational response generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 270–278 (2020)
145. Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X.: Personalized dialogue generation with diversified traits. arXiv preprint arXiv:1901.09672 (2019)
146. Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X.: Personalized dialogue generation with diversified traits (2020)
147. Zheng, Y., Zhang, R., Huang, M., Mao, X.: A pre-training based personalized dialogue generation model with persona-sparse data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9693–9700 (2020)
148. Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: emotional conversation generation with internal and external memory. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
149. Zhou, X., et al.: Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1118–1127 (2018)
150. Zukerman, I., Litman, D.: Natural language processing and user modeling: synergies and limitations. User Model. User-Adap. Interact. **11**(1), 129–158 (2001). https://doi.org/10.1023/A:1011174108613