

A Methodology for Evaluating Interaction Strategies of Task-Oriented Conversational Agents

Marco Guerin^{1,2}, Sara Falcone¹, Bernardo Magnini¹

¹Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento — Italy

²AdeptMind Scholar, Canada

{guerini, sfalcone, magnini}@fbk.eu

Abstract

In task-oriented conversational agents, more attention has been usually devoted to assessing task effectiveness, rather than to *how* the task is achieved. However, conversational agents are moving towards more complex and human-like interaction capabilities (e.g. the ability to use a formal/informal register, to show an empathetic behavior), for which standard evaluation methodologies may not suffice. In this paper, we provide a novel methodology to assess - in a completely controlled way - the impact on the quality of experience of agent's interaction strategies. The methodology is based on a within subject design, where two slightly different transcripts of the same interaction with a conversational agent are presented to the user. Through a series of pilot experiments we prove that this methodology allows fast and cheap experimentation/evaluation, focusing on aspects that are overlooked by current methods.

1 Introduction

The evaluation of task-oriented conversational agents is usually focused on measuring their effectiveness, either at the single turn level - see for example (Wen et al., 2015; Frampton and Lemon, 2006; Chen et al., 2013) - or at the level of the whole interaction - e.g. success rate (Dybkaer et al., 2004). Still, as conversational agents are becoming more complex and human-like (Bowden et al., 2017; Romero et al., 2017; Cercas Curry et al., 2017), these evaluation methodologies may not suffice. In this paper, we present a framework for evaluating interaction strategies of conversational agents during their development phase. Our approach combines in a novel way methodologies already tested and validated, and is based on

Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, 978-1-948087-75-9

a pairwise comparison of manually curated transcripts of possible interactions.

On the one hand, our methodology is inspired by the Human-Computer-Interaction (HCI) literature by dividing the evaluation of a system in the *Quality of Service* (QoS) and *Quality of Experience* (QoE) dimensions (Moller et al., 2009). The former corresponds to the efficiency of the system, while the latter refers to the way in which the system accomplishes the task. In dialogue systems evaluation the traditional focus is on the QoS, while in this work we deal also with the QoE. On the other hand, we take advantage of crowdsourcing methodologies, a fast and cheap way we use to evaluate interactions while maintaining complete control over experimental conditions – by using a design similar to A/B testing, but in a ‘within subject’ condition. In this setting two slightly different versions of the same interaction with a conversational agent are presented to the user for a pairwise comparison (e.g. the same interaction using a formal/informal register). Unlike standard Wizard of Oz (WoZ) or lab experiments, the user does not directly interact with the system, rather s/he reads the manually curated transcript, so to eliminate confounding variables and make data collection much faster.

The paper is structured as follows: in Section 2 we discuss some of the main approaches used in the evaluation of conversational agents. In Section 3 and 4 we present our framework and provide some pilot experiments respectively. Finally in Section 5 we discuss the advantages of the approach in light of the results of the experiments.

2 Related Works

Several frameworks to evaluate dialogue systems have been proposed. So far, evaluation mainly focused on implemented components/systems and

followed different criteria taken from other research fields, such as machine translation (Wen et al., 2016), human-computer interaction (Allen et al., 2001), user experience and interfaces design (Skantze, 2005). The fact that these methodologies are not designed to evaluate dialogue system, can affect the results - for example, machine translation metrics do not correlate well with human judgments (Liu et al., 2016). Another common aspect of these approaches is that they rely on a complete implementation of the system to evaluate aspects such as efficiency (Raux et al., 2006), quality (Shawar and Atwell, 2007) or both (Silvervarg and Jönsson, 2011), while in the case of our interaction strategies it would be useful to have a simulation approach that allows to predict the possible impact of such strategies. In the following we first discuss standard methodologies for implemented systems, then methodologies using simulation, and finally evaluation in related fields that inspired our approach.

Evaluation of implemented systems. Among the metrics used for evaluating specific components of a system we can briefly mention: (i) *fluency/grammaticality* of the generated sentences in the NLG step of the interaction, that can be done either manually (Wen et al., 2015) or in a semi-automatic way, as in (Riezler et al., 2003); (ii) *slots correctly realized*, an automatic evaluation of the NLG component (Scheffler and Young, 2002; Frampton and Lemon, 2006); (iii) *slots correctly recognized*, an automatic technique used to evaluate the NLU component (Levin and Pieraccini, 1997; Chen et al., 2013).

Among the metrics used for evaluating whole interactions there is *success rate*. It can be based on objective automatic measures or on a subjective evaluation made by users evaluating the system according to guidelines provided by the experimenter (Dybkjaer et al., 2004).

Finally, a framework worth mentioning is PARADISE (Walker et al., 1997) that is specifically devoted to spoken dialogue systems (while in our work we consider text based interactions only). This work focuses on metrics such as task success rate and dialogue cost (e.g. dialogue time, number of utterances, agent response delay) to evaluate the quality of a system. With regard to spoken dialogue systems, the use of crowdsourcing for collecting preference judgments has already been explored, for example in (Trippas et al., 2017; Chuk-

lin et al., 2018; Alfonseca, 2017)

Evaluation through simulation. If the system is still at an early stage of development, a viable solution is to use *WoZ experiments* (Dahlbäck et al., 1993; Paek, 2001; Raux et al., 2006), in which the interaction is simulated and users are prepared on how to behave. Still, this approach suffers of some main drawbacks: (i) the need for conducting several time-consuming interactions to get stable results; (ii) the possible measured improvements of the system can still be biased by confounding variables; (iii) it is difficult for wizards to provide consistent responses across sessions; (iv) ‘behavior instructions’ should be prepared and given to the wizard and possibly to each single user¹ (v) these ‘behavior instructions’ cannot describe every single reaction, but must try to control typical situations.

Evaluation in related fields. Our design leverages in a novel way elements used in several fields.

Two variants testing with controlled stimulus material. In the MT field, the work by (Graham et al., 2013) used a ‘within subject’ design where each evaluator was sometimes presented with a small random textual variation (control condition) of a translation they were already exposed to (experimental condition). This methodology was used to evaluate the quality of raters’ judgments. Closely to our approach, the MT evaluation campaign presented in (Bojar et al., 2016) used expert annotators for *pairwise* system comparisons denoting whether a system A was judged better than, worse than, or equivalent to another system B. In this case the two conditions were presented simultaneously, side by side, rather than in a random sequential order as in (Graham et al., 2013). Other seminal approaches - using direct comparison of stimulus materials via pairwise comparison - is presented in the realm of affective NLG (Van Der Sluis and Mellish, 2010), and in the domain of persuasive NLP (Tan et al., 2014). Still, both works used this procedure just for the validation of stimulus material and made resort to traditional evaluation procedures for the final evaluation. Finally, in the realm of persuasive NLG a crowdsourced approach based on A/B testing

¹ e.g. ‘pretend you are sad because ...’ so to trigger the desired system response, such as empathy. In fact, if the user were totally ‘free’ to interact with the Wizard s/he could miss the functionality under inspection – Still, guiding the user during the interaction strongly affects its naturalness. On the other hand Wizards require significant training so to respond in a way that is credible and consistent.

and focused on ecological validity is presented in (Guerini et al., 2012). This approach, however, uses a between-subject design, where subjects are presented with just one stimulus material.

Transcripts and ‘third party’ evaluation. Two approaches that use transcripts of the conversation, instead of a direct interaction with the agent, are presented in (Jurčiček et al., 2011; Yang et al., 2010). These works compared lab experiments with crowdsourced ones - in the scenario of spoken dialogue systems - showing that the results in the former (direct interaction with the system) are comparable with the results in the latter (third party users reading transcriptions). Similarly, in (Pragst et al., 2017), the authors focus on a WoZ evaluation of the interaction strategies of an embodied conversational agents. Users were presented with the video of an embodied conversational agent interacting with a human user (the agent was guided by a Wizard and the user was an instructed actor). The subjects have to evaluate the interactions answering to a survey using a Likert scale. In this experiment, as in the previous one, the subject is third-party evaluator who did not directly entered the interaction.

3 Proposed Solution

Starting from the advantages and limitations of the previous approaches, we designed a new framework to evaluate a task-oriented dialogue system from the point of view of the strategies of interaction. In our framework the dependent variables are QoS and QoE aspects instantiated in a questionnaire to be evaluated by the subjects, while the independent variables are the interaction strategies that are instantiated in the stimulus material.

In particular, we propose a methodology in which the transcripts of two versions of the interaction with a conversational agent (e.g. one using a formal language and one using an informal one, one being empathetic and one not) are presented to the user, to see if one version is preferred over the other. The core idea of the approach is that, differently from WoZ studies, the subjects must *read* the transcripts of the interaction rather than directly *interacting* with the agent. This is required in order to grant complete control over the experiment (transcripts can be manually curated so to meet stringent control criteria). The two versions must maintain all aspects and wording of the interaction the same (apart from those affected by the modal-

ity being tested), including the outcome (e.g. success of the interaction) so that, if one version is preferred over the other, we can conclude that the effect of preference is solely due to the variable of interest (e.g. the “formality level” of the language, the empathy of the agent) and not to other factors.

The procedure for setting up an experiment is:

1. *Control conditions.* Create one or more control conditions for each interaction strategy to be tested: either a transcript of a real interaction with an existing system or a possible interaction with the planned one.
2. *Experimental conditions.* Create an experimental condition that is the manually curated counterpart of the control condition. As stated, changes in the wording should be minimal and must always reflect the interaction strategy to be tested. Changes can be of two types: (a) *substitution* of portions of system’s utterances with new coherent portions that represent the experimental condition (e.g. change an informal greeting with a formal one) or (b) *insertion* of new portions of text in system’s utterances.
3. *Questionnaire.* Prepare a questionnaire that includes questions about the QoS and QoE dimensions of interest.
4. *Crowdsourcing.* Built a task on a crowdsourcing platform with a pairwise comparison design and the questionnaire subministered after each comparison.

Many interaction strategies can be analysed to test our approach. We decided to focus on five of them, those we deemed most interesting and impactful on the pragmatics of the dialogue and for which an effect should be detected (Radziwill and Benton, 2017), so to test if our methodology is able to capture such effect.

4 Experiments

In this section we describe a showcase experiment for our methodology, where we evaluated 5 possible variants of CH1, a conversational agent that we implemented in order to calculate the carbohydrates of user’s meals. We set up a two variants testing for each independent variable, where we provided to the subjects of the experiment the transcripts of some conversations between a human user and CH1. Before starting the experiment, the user received a short text describing the task.

4.1 Interaction Strategies

Five strategies, together with their linguistic parameters, were analyzed. The transcripts of the experimental condition were realized by two expert linguists, following the substitution/insertion instructions described in Section 3.

Empathy can be defined as the ability of a conversational agent to adapt to the user feelings and also to provide flexible emotionally-coloured responses for different purposes (Callejas et al., 2011). There exist many different ways in which emotions are defined, represented and managed within dialogue systems (Meira and Canuto, 2015; Barrett et al., 2007). Usually, the recognition is based on the manifestation of the user emotion, which can be processed considering linguistic (Balahur et al., 2014) and paralinguistic cues (Schuller et al., 2013).

Formality in linguistics is expressed through the choice of lexical expressions. According to the context, the speaker can use a specific linguistic register, style and lexicon (Heylighen and Dewaele, 1999). In order to detect the formality of a text there exist different strategies. One is to detect the average of deixis for each grammatical category of words (Heylighen and Dewaele, 1999); another is to use words length and latinate affix (Brooke et al., 2010).

Facing is the ability to tackle situations in which the conversational agent has not a proper or pre-set answer (Morrissey and Kirakowski, 2013). We can observe two kinds of facing for unexpected users' input: (i) the agent is not able to recognize the intention and makes resort to a default answer, e.g. "Sorry I do not understand, could you repeat?"; (ii) the agent is able to recognize the intention and it provides a suitable/contextual answer even if it is not endowed with the skills to solve it.

Vocabulary Extension concerns agent's ability to learn new words during the conversation and use them appropriately in the ongoing (Riccardi and Hakkani-Tur, 2005). For example, CH1 needs to know a huge variety of food names (from specific names such as 'seitan' to complex recipes such as 'plantain coated sea bass with mango wine sauce') to calculate meals carbohydrates. Therefore, since covering all possible combinations of ingredients and recipes is almost impossible, the ability to learn new food names during the interaction improves user experience.

Linguistic Alignment corresponds to the con-

versational agent functionality of adapting its language to that of the user. The agent will start using the user's frequent expressions in order to align its lexicon. For example, it should align its linguistic register or reuse the same words used by the user in the generation of the following turn (Branigan et al., 2010; Duplessis et al., 2017).

In Table 1 we give, as an example, the transcript used as stimulus material for the *empathy* variable.

4.2 Dependent Variables

The variables that we adopted in our framework for evaluating QoS and QoE are: (i) **utility**: if the user found the system useful to achieve the task and obtained all the information s/he needed; (ii) **ease of use**: if the system was intuitive in the usage and the user could use it without effort; (iii) **satisfaction**: if the user had a good experience and would use the system again; (iv) **interaction**: if the user appreciated the manner of interacting of the system. The evaluation of these variables has been obtained asking the subjects to choose the interaction that better matched each of the four questions under each interaction pair. According to the kind of system that has to be evaluated, different or more fine grained dependent variables can be chosen. For example, the *cognitive workload* or *effort* perceived by the user, the *appeal* of the interface design or the communication channel.

4.3 Experiment description

In this section we describe the main characteristics of our evaluation experiment.

Subjects: 143 subjects from the US were recruited using the CrowdFlower platform: 93 male and 50 female. 36 were between 18-24 years old, 58 were between 25-34 years old, 31 were between 35-49 years old, 18 were 50 or more aged.

Design: The design was completely within-subject, i.e. each subject was presented with one of the control and experimental transcripts for the 5 variables. Transcripts order among variables and between control/experimental conditions was randomized in order to avoid any framing effect or stimulus order effect (Kessler and Meier, 2014).

Quality control: all subjects were level 3 contributors (maximum expertise/reliability) and a minimum of 3 minutes was set to accept the responses to the questionnaire. No "gold-standard" item was used to evaluate rater reliability, as the two former controls proved to be enough for our case, as found in post hoc analysis.

<p>CH1: Hello Andrea! What did you eat for your last meal? User: I ate a plate of spaghetti with tomato CH1: How much spaghetti with tomato did you eat? User: Unfortunately I messed up, it was a generous helping... CH1: Your meal consisting of a generous plate (200g) of spaghetti with tomato corresponds to 30.85 g of carbohydrates. I hope I have been helpful! See you soon!</p>	<p>CH1: Hello Andrea! What did you eat for your last meal? User: I ate a plate of spaghetti with tomato CH1: How much spaghetti with tomato did you eat? User: Unfortunately I messed up, it was a generous helping... CH1: Don't worry about it, everyone messes up sometimes! Anyway, your meal consisting of a generous plate (200g) of spaghetti with tomato corresponds to 30.85 g of carbohydrates. I hope I have been helpful! See you soon!</p>
---	--

Table 1: Control (on the left) and experimental (on the right) transcript for the *empathy* independent variable. Portions of CH1 utterances that were changed in order to realize the variable are in bold.

	Ease	Satisf.	Util.	Inter.	Marginal
alignment	0.60	0.61	0.67	0.65	0.63
empathy	0.73	0.78	0.73	0.76	0.75
facing	0.64	0.71	0.70	0.66	0.68
formal	0.74	0.80	0.73	0.66	0.73
vocabulary	0.74	0.71	0.73	0.77	0.74

Table 2: Ratio of subjects that preferred the experimental over the control condition.

Judgments collected: the total number of judgments collected is 2860: 143 subjects that answered four questions for each of the 5 independent variables.

Cost: Overall, the experiment cost was 51.48\$ resulting in a cost of roughly 10\$ for evaluating each variable. The duration of the experiment was about 12 hours. As a side note, the experiment got a high feedback in terms of contributor satisfaction (an overall evaluation of 4.8/5).

4.4 Results

In this section we briefly discuss the results, reported in Table 2, of our pilot experiments. We focus on the ability of our framework to elicit in users' responses a difference between the two levels of each independent variable in terms of perceived QoS and QoE. Results were in line with our expectations: the methodology was able to capture the effect of each modality and strategy of interaction in the experimental condition.

Results shows, indeed, that the contributors expressed a preference for the experimental condition, resulting in a consistent trend with respect to the variables². All results are statistically significant, χ^2 test used. Moreover, the independent variables have different magnitude effects (i.e. some

²Actually, for the formal/informal dimension the preference went to the control condition (formal register). Still, for comparability purposes we report results for the control condition in Table 2

	FEMALE	MALE
alignment	0.68	0.61
empathy	0.77	0.74
facing	0.76	0.64
formal	0.76	0.72
vocabulary	0.72	0.75

Table 3: Marginals for the interaction variables according to gender.

modalities of interaction were appreciated more). In particular, considering marginals, *empathy*, *formality* and *vocabulary* were the most appreciated variations of CH1 (with no statistical significant difference among them) while *alignment* and *facing* were less appreciated. Interestingly, an analysis at the gender level (see Table 3), revealed that on the two latter variables there was a clear discrepancy in the marginals between male and female: this difference in the case of *alignment* is 0.68 for female vs. 0.61 for male - and both account for the difference in overall results with regard to other independent variables. Instead, for *facing*, the difference in marginals with regard to other independent variables was due to the male group alone, since for female the results are in line with other variables (0.64 vs. 0.76).

Turning to dependent variables we can see that the effect is quite different: *alignment* has a main impact on utility and interaction, *empathy* on satisfaction and interaction, *facing* on satisfaction and utility, *formality* on satisfaction and ease of use, *vocabulary* on naturalness and ease of use. Interestingly each of the independent variables had a main effect on one QoS and one QoE dimension - in line with the findings of (Jurčiček et al., 2011).

4.5 Comparison with WoZ

Finally, we simulated a WoZ experiment in order to compare the design, implementation and performance of our framework. While the instruction

and stimuli creation require in both cases almost the same time (for example the stimulus material for our setting was used as an example of possible interaction for the Wizard instructions), the implementation of our framework is much faster. Indeed, the WoZ experiment requires the implementation of a graphical user interface, but even if we use a pre-set one, we still need to instruct Wizard(s) and find a relevant number of participants in case a crowdsourcing methodology is not used. But even if we do not consider the aforementioned time consuming preparatory activities, each WoZ session that replicate our experiment, required 30 minutes and two participants, as compared to the 3 minutes and one participant required by our framework. This is explained by the fact that while in our framework the subject just need to read the transcript of the interaction, in the WoZ experiment the user needs to read instructions for each interaction, think and digit the input at each turn and read the corresponding wizard response; at the same time the Wizard needs to do the same.

5 Advantages

With the initial evidence, provided by the experiments, we can reasonably state that the framework we are proposing has some important advantages:

Cheap and Fast. The evaluation can be obtained using platform such as CrowdFlower or AMT, choosing high level and possibly native speaker contributors. Crowdsourcing approaches make it quick and cheap to run evaluation experiments as compared to ecological ones, see for example what reported in (Reiter, 2011).

Flexibility. The framework gives the possibility to define the dependent and independent variables that better match the strategies and modalities of interaction that need to be evaluated. Moreover, using crowdsourcing approaches together with hand curated transcripts we can easily experiment several variables/versions of the conversational agents or control for multiple mixed effects (e.g. linguistic style * empathy). We can also test different levels of a strategy, for example to find the optimal formality level.

Experiment design. the adoption of a pairwise comparison of the two versions of the system makes the evaluation of the interaction strategies faster and more direct. It also halves the number of judgments required with respect to traditional evaluation designs in which each stimulus mate-

rial is served separately, bringing to an approximate halving of the price.

Control over the variables being tested. Providing transcripts of the conversation to the subjects gives the possibility to control one variable at a time isolating its effect (and to the best of our knowledge no previous work ever tried this approach). This allow us, for example, to build transcripts with an almost equal number of tokens and turns of interactions, in order to avoid phenomena such as length effect (Koizumi, 2012).

Judgement Elicitation. Forcing a choice between control and experimental condition allows eliciting possible differences between the two interactions, for how small this difference could be.

Effort Reduction. Since the subjects of the experiment are not meant to interact directly with the conversational agent, we can create an off-line experiment to test conversational agents characteristics in advance, rather than having a post-process analysis. This saves implementation or data collection effort, since there might be aspects of the interaction that annoy the user or, on the contrary, that have a positive impact and that are easy to implement. Finally, we can avoid the risk that the user could miss some passages of the interaction useful to highlight the strategies that we are analyzing, as could happen in WoZ studies.

6 Conclusion and Future Works

In our view, the proposed framework, based on a pairwise comparison of manually curated and controlled transcripts, represents a step forward in the evaluation of dialogue systems. This methodology allows evaluating the strategies and the interaction modalities of a conversational agent before its implementation, ensuring the advantages reported above. We believe that this methodology is suitable not only for rule-based systems, but also for data-driven ones. In this latter case the methodology can be used, for example, to define the constraints for data collection.

In future works, we would like to define and test other strategies of interaction, but it might be necessary - to create proper transcripts - to define new guidelines and parameters. For example if a strategy involves choosing between two different dialog paths (i.e. several turns might change) the guidelines on insertion or substitution we defined are not sufficient.

References

- Enrique Alfonseca. 2017. Evaluation of speech for the google assistant. <https://ai.googleblog.com/2017/12/evaluation-of-speech-for-google.html>.
- James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI magazine*, 22(4):27.
- Alexandra Balahur, Rada Mihalcea, and Andrés Montoyo. 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications.
- Lisa Feldman Barrett, Kristen A Lindquist, and Maria Gendron. 2007. Language as context for the perception of emotion. *Trends in cognitive sciences*, 11(8):327–332.
- Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16)*, pages 131–198. The Association for Computational Linguistics.
- Kevin K Bowden, Shereen Oraby, Amita Misra, Jiaqi Wu, and Stephanie Lukin. 2017. Data-driven dialogue systems for social agents. *arXiv preprint arXiv:1709.03190*.
- Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.
- Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 90–98. Association for Computational Linguistics.
- Zoraida Callejas, David Griol, and Ramón López-Cózar. 2011. Predicting user mental states in spoken dialogue systems. *EURASIP Journal on Advances in Signal Processing*, 2011(1):6.
- Amanda Cercas Curry, Helen Hastie, and Verena Rieser. 2017. A review of evaluation techniques for social dialogue systems. *arXiv preprint arXiv:1709.04409*.
- Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 120–125. IEEE.
- Aleksandr Chuklin, Aliaksei Severyn, Johanne Trippas, Enrique Alfonseca, Hanna Silen, and Damiano Spina. 2018. Prosody modifications for question-answering in voice-only settings. *arXiv preprint arXiv:1806.03957*.
- Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies - why and how. *Knowledge-based systems*, 6(4):258–266.
- Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81.
- Laila Dybkjaer, Niels Ole Bernsen, and Wolfgang Minker. 2004. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54.
- Matthew Frampton and Oliver Lemon. 2006. Learning more effective dialogue strategies using limited dialogue move features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 185–192. Association for Computational Linguistics.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.
- Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2012. Ecological evaluation of persuasive messages using google adwords. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 988–996. Association for Computational Linguistics.
- Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants.
- Filip Jurčićek, Simon Keizer, Milica Gašić, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2011. Real user evaluation of spoken dialogue systems using amazon mechanical turk. In *Twelfth Annual Conference of the International Speech Communication Association*.
- Judd B Kessler and Stephan Meier. 2014. Learning from (failed) replications: Cognitive load manipulations and charitable giving. *Journal of Economic Behavior & Organization*, 102:10–13.
- Rie Koizumi. 2012. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, 1(1):60–69.
- Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Fifth European Conference on Speech Communication and Technology*.

- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- MO Meira and AMP Canuto. 2015. Evaluation of emotional agents’ architectures: an approach based on quality metrics and the influence of emotions on users. In *Proceedings of the World Congress on Engineering*, volume 1.
- Sebastian Moller, Klaus-Peter Engelbrecht, Christine Kuhnel, Ina Wechsung, and Benjamin Weiss. 2009. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Quality of Multimedia Experience, 2009. QoMEX 2009. International Workshop on*, pages 7–12. IEEE.
- Kellie Morrissey and Jurek Kirakowski. 2013. In *International Conference on Human-Computer Interaction*, pages 87–96. Springer.
- Tim Paek. 2001. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, page 2. Association for Computational Linguistics.
- Louisa Pragst, Wolfgang Minker, and Stefan Ultes. 2017. Exploring the applicability of elaborateness and indirectness in dialogue management.
- Nicole M Radziwill and Morgan C Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.
- Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of let’s go! experience. In *Ninth International Conference on Spoken Language Processing*.
- Ehud Reiter. 2011. Task-based evaluation of nlg systems: Control vs real-world context. In *Proceedings of the UCNLG+ Eval: Language Generation and Evaluation Workshop*, pages 28–32. Association for Computational Linguistics.
- Giuseppe Riccardi and Dilek Hakkani-Tur. 2005. Active learning: Theory and applications to automatic speech recognition. *IEEE transactions on speech and audio processing*, 13(4):504–511.
- Stefan Riezler, Tracy H King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 118–125. Association for Computational Linguistics.
- Oscar J Romero, Ran Zhao, and Justine Cassell. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3807–3813. AAAI Press.
- Konrad Scheffler and Steve Young. 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of the second international conference on Human Language Technology Research*, pages 12–19. Morgan Kaufmann Publishers Inc.
- Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.
- Bayan Abu Shawar and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 89–96. Association for Computational Linguistics.
- Annika Silfvervarg and Arne Jönsson. 2011. Subjective and objective evaluation of conversational agents in learning environments for young teenagers. In *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.
- Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341.
- Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*.
- Johanne R Trippas, Damiano Spina, Lawrence Cave-don, and Mark Sanderson. 2017. Crowdsourcing user preferences and ery judgments for speech-only search. In *1st SIGIR Workshop on Conversational Approaches to Information Retrieval (CAIR’17)*.
- Ielka Van Der Sluis and Chris Mellish. 2010. Towards empirical evaluation of affective tactical nlg. In *Empirical methods in natural language generation*, pages 242–263. Springer.
- Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Zhaojun Yang, Baichuan Li, Yi Zhu, Irwin King, Gina Levow, and Helen Meng. 2010. Collection of user judgments on spoken dialog system with crowdsourcing. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 277–282. IEEE.