Overview of the EVALITA 2018 Evaluation of Italian DIALogue systems (IDIAL) Task

Francesco Cutugno[1], **Maria Di Maro**[1], **Sara Falcone**[2,3],
Marco Guerini[2], **Bernardo Magnini**[2], **Antonio Origlia**[1]
[1] Università degli Studi di Napoli 'Federico II'
[2] Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento — Italy
[3] University of Trento, Italy.
{cutugno, maria.dimaro2, antonio.origlia}@unina.it
{sfalcone, guerini, magnini}@fbk.eu

Abstract

English. We report about the organization of the IDIAL (Evaluation of Italian DIALogue systems) task at EVALITA 2018, the first shared task aiming at assessing interactive characteristics of conversational agents for the Italian language. In this perspective, IDIAL considers a dialogue system as a "black box" (i.e., evaluation can not access internal components of the system), and measures the system performance on three dimensions: task completion, effectiveness of the dialogue and user satisfaction. We describe the IDIAL evaluation protocol, and show how it has been applied to the three participating systems. Finally, we briefly discuss current limitations and future improvements of the IDIAL methodology.

Italiano. *Riportiamo circa l'organizzazione del task IDIAL (Valutazione di sistemi di dialogo per l'italiano) a Evalita 2018. IDIAL é il primo task condiviso per la valutazione delle caratteristiche di interazione di agenti conversazionali per l'italiano. In questa prospettiva, IDIAL considera un sistema di dialogo come una "black box" (in quanto la valutazione non puó accedere ai componenti interni del sistema), e misura le prestazioni del sistema su tre dimensioni: la capacitá di portare a termine il task, l'efficacia del dialogo, e la soddisfazione dell'utente. Descriviamo il protocollo di valutazione IDIAL, e mostriamo come esso é stato applicato a tre sistemi partecipanti. Infine, discutiamo brevemente le limitazioni attuali e i miglioramenti futuri della metodologia.*

1 Task Motivations

The IDIAL (Evaluation of Italian DIALogue systems) task at EVALITA 2018 (Caselli et al., 2018) intends to develop and apply evaluation protocols for the quality assessment of existing task-oriented dialogue systems for the Italian language. Conversational Agents are one of the most impressive evidence of the recent resurgence of Artificial Intelligence. In fact, there is now a high expectation for a new generation of dialogue systems that can naturally interact and assist humans in several scenarios, including virtual coaches, personal assistants and automatic help desks. However, despite the growing commercial interest for various task-oriented conversational agents, there is still a general lack of methodologies for their evaluation. During the last years, the scientific community has studied the evaluation of dialogue systems under different perspectives, concerning for instance the appropriateness of the answer (Tao et al., 2017), or user satisfaction metrics (Hartikainen et al., 2004; Guerini et al., 2018).

IDIAL proposes an objective evaluation framework, which consists in both of user perception towards the ease of use and the utility of the system, and of consistency, robustness and correctness of the task-oriented conversational agent. The IDIAL starting point are previous evaluation models, comprising the observation of users and systems' behaviour, the judgment process inside the users, and the quality of the system regarding its service objectives (Möller and Ward, 2008).

2 IDIAL Evaluation Protocol

IDIAL assumes that the systems under assessment are task-oriented dialogue systems (TODSs), providing specific services in an application domain, such as hotel booking or technical support service. Systems can be either monomodal (spoken or written) or multimodal, and are used for completing a number of predefined tasks (or intents), each of which can be achieved in one or more interactions, or conversational turn pairs (i.e. sometimes a question-answer pair is not sufficient to accomplish the intended action, since other conversational turns are needed). TODS to be examined can be on-line or off-line applications, and are evaluated as "black-boxes", meaning that evaluators will not have access to the internal characteristics and components of the system. Given the peculiar nature of the evaluation, which is carried out by human users, IDIAL does not require neither training nor testing data. We target the evaluation of existing TODSs (both industrial and academic prototypes), which are on operation at the date of the test period (September 2018). The output of the evaluation is not a ranking. Conversely, we provide a qualitative assessment for each participating system, based on detailed and coherent set of technological and interactive characteristics of the system.

2.1 Evaluation Method

The IDIAL evaluation procedure is defined to address the following three characteristics of a task oriented-conversational agent:

A. Task completion. This is the capacity of the system to achieve the goals of the task for which the system has been designed, in a reasonable amount of time.

B. Effectiveness of the dialogue. This is the capacity of the system to interact with the user in order to achieve its task. It includes, among the others, the capacity to interpret commands accurately, the robustness of the system to unexpected input, the ease of use of the system, and the fluency of the dialogue.

C. User satisfaction. This is the reaction of the user after having used the system. It includes aspects like the degree of empathy of the system, the ability to read and respond to moods of human participant, the capacity of the system to give conversational cues, and the use appropriate degrees of formality.

The three characteristics (A-C) mentioned above are assessed in IDIAL by means of two evaluation methods, a questionnaire, and a set of linguistic stress tests.

Questionnaire. A questionnaire is given to the user after s/he has interacted with the system for a certain number of tasks. Questions may address each of the three main behaviours of the system (task completion, effectiveness of the dialogue and user satisfaction), and require the user to estimate the degree of acceptability (on a Likert scale), of a number of statements about the system. The questionnaire is prepared by experts, and it is intended to address questions both about Quality of Service and Quality of Experience. Whereas Quality of Service is about the accomplishment of the task, concerning the correct transferring of the needed information to the user, Quality of Experience consists of how the task was accomplished, if the user enjoyed the experience and would use the system again or recommend it (Moller et al., 2009).

Linguistic stress tests. A stress test is intended to assess the system behaviour under an unconventional interaction situation (i.e. a stressful situation), in order to evaluate the robustness of the system itself. In IDIAL 2018 we consider only linguistic stress tests, which are designed and applied by expert computational linguists. Stress tests are applied on real interactions through a substitution mechanism: given a user utterance in a dialogue, the utterance is minimally modified substituting some of the elements of the sentence, according to a pre-defined list of linguistic phenomena (e.g. typos, lexical choices, different kinds of syntactic structures, semantic reformulations, anaphora resolution). Other phenomena related to dialogue (e.g. requests of explanation, requests of interruption of the conversation) have not be considered in IDIAL 2018, and will be discussed for future editions. After application, a stress test is considered as "passed" if the behaviour of the system is not negatively affected by the substitution, otherwise the application is considered as "fail". The final score for a system is given by the ratio between the number of successfully applied stress tests over the total number of applied stress tests.

Table 1 summarize the system behaviours that are considered by the IDIAL evaluation, as well as the respective evaluation tools and their expected output.

2.2 Evaluation Procedure

Given a dialogue system to be evaluated, the evaluation phases described in Table 1 are practically applied according to the following steps:

1. Organizers prepare a user satisfaction ques-

System behaviour	Evaluation tool	Evaluation output
A. Task completion	Questionnaire	Summary report based on average scores on Likert scale
B. Effectiveness of the dialogue	Stress tests + Questionnaire	Summary report based on stress test success rate + summary report based on average scores on Likert scale
C. User satisfaction	Questionnaire	Summary report based on average scores on Likert scale

Table 1: Summary of IDIAL evaluation protocol.

tionnaire, which will be applied to all systems under evaluation (i.e. the questionnaire is not personalized). The questionnaire is reported in Section 4, and the Italian version is available as Appendix A of the IDIAL evaluation protocol.

2. Organizers write instructions on how to use the system on the base of a system submission (see Section 2.3). Typical instructions contain a task to be achieved by using the system (e.g. "book a train ticket for one person – you are free to decide destination and date").

3. Organizers individuate few users with average expertise for the system domain and task. For instance, if the system has been designed to serve a high school student, it seems reasonable to involve high school students as users.

4. Selected users interact with the system in order to achieve the goals defined at point 2. All interactions are recorded, and logs are made available. Depending of the task complexity, organizers decide how many runs will be experimented with the system. Overall, each user should not spend more than one hour with a system.

5. Just after the interaction, organizers provide each user with the questionnaire to be filled in. The same questionnaire is used for all participating systems.

6. Organizers select a sample of user interactions, and use them to design applicable stress test. The stress tests actually implemented in the evaluation are reported in Section 4.

7. Organizers run stress tests on user interactions and record system behaviour. In order to keep the experimental setting under control, only one stress test per interaction is applied.

8. For each system, organizers write the final evaluation summary report, on the base of both the questionnaire and the stress tests, according to the metrics reported in Table 1.

2.3 Submission Requirements

At submission time, IDIAL participants are asked to provide the following information concerning their system.

- Specify the tasks that the system can do, in the form of user intents (e.g. buy a train ticket, search for point of interest, take an appointment for a meeting, block credit card, etc.). More than one task for a system are allowed.

- Specify as much as possible the application domain of the system, in order to understand the knowledge which is managed during a dialogue (e.g. Italian railway stations, restaurants in Trento, meetings within one month, etc.).

- Interaction channel of the system (i.e. spoken, written, multimodal).

- System interface (e.g. messenger, telegram, twitter, proprietary interface, etc.).

- Access to the system (i.e. on-line, off-line, telephony service, etc.).

3 Participant Systems

Three different dialogue systems were tested according with the IDIAL evaluation protocol: a chatbot developed by the NLP research group at *Fondazione Bruno Kessler*[1], aiming at calculating the amount of carbohydrates in a meal; a vocal call-steering service developed by the Italian company *Interactive Media*[2], already in operation at a financial company, aiming at understanding the customer call and routing the call either to a human operator or to an automatic service; a spoken dialogue system developed by the Italian speech recognition company *Cedat 85*[3], aiming at supporting customers of a telco company in a number of services. The three systems, being the result of the research of groups with diverse background and application goals, were therefore very different in nature. Hence, this allowed us to test the *scalability* of the proposed protocol.

3.1 CH1 Conversational System for Diabetics

CH1 is a prototype (i.e. it is not yet operative) conversational agent capable of computing the grams of carbohydrates in a meal (Magnini et al., 2018). The chatbot is based on a written interaction in Italian, runs on Telegram and is designed to help diabetics who need to perform a "carbs count" for each consumed meal. The interaction is system-initiative, starting with a question posed by the bot concerning with the food eaten by the users during their last meal. The conversational exchange goes on with the list of food given by the user. In case the typed keywords do not exactly correspond to the vocabulary known by the system, the system provides a list of most similar dishes or ingredients to correctly compute the quantity of carbohydrates. The knowledge base used both to extract the similar food and to perform the carbohydrates computation is a domain ontology called Hellis, based on available nutritional scientific literature. The system makes use of machine learning approaches trained on a manually-annotated Italian corpus (DPD - Diabetic Patients Diary), containing diary entries written by diabetic patients.

3.2 Interactive Media Call-Steering System

Interactive Media submitted to IDIAL a virtual agent for receiving and routing calls to human or automatic service operators. The system is operative at the time of IDIAL evaluation. The spoken interaction takes place via the telephone channel and is user-initiative. As a matter of fact, after the initial user identification, through which the system asks name, surname and date of birth of users for their recognition, the human interlocutor is left free of asking open questions related to the field of application of the system itself, i.e. customer service for banks (for instance, *I want to get a loan*), call-steering for offices and companies, etc. Afterwards, the system can ask questions for disambiguation purposes (for instance, *Are you interested in a loan higher or lower than 3000?*), in order to properly classify the call in the correct category for the proper operator. The system has been developed within the IM-MIND platform integrated with Cisco CTI (*Computer Telephony Integration*)[4] and Nuance speech technologies[5].

3.3 Cedat 85: Speech Technologies in Action

The spoken dialogue system provided by Cedat 85 is a prototype (i.e. it is not yet operative) performing specific tasks suggested by the system at the beginning of the interaction (system-initiative) at the telephone. The following tasks can be addressed: i) informative operations concerning the final invoice, the list of transactions, the phone credit, and the tariff plan; ii) active operations such as loading your prepaid phone card (specifying the amount of money) and making a wire transfer (specifying the amount of money and the recipient). In case the request is not well understood, the system guides the user clarifying the possible actions to be performed. After that the user intent is understood and carried out, the user can continue with a new request or end the conversation.

4 Application of the IDIAL Evaluation protocol

The evaluation of the participating dialogue systems, as introduced in Section 2, was accomplished through two modules:

1. User experience, measured through a questionnaire;

2. Stress tests, mostly based on the log of the previous evaluation. The stress tests were evaluated using a pass/fail modality.

[1] https://ict.fbk.eu/units/nlp/

[2] https://www.imnet.com/it/

[3] http://www.cedat85.com/

[4] https://www.cisco.com/

[5] https://www.nuance.com/index.html

4.1 User Experience

We selected 10 different users for each system, who differed for age (range 19-60), sex and cultural background, for a total of 30 users. Each user had to interact with the system for three randomized tasks, for a total of 30 interactions for a system. After the completion of the tasks, each user was asked to fill in a questionnaire. It took in average 10 minutes to explain the tasks to be accomplished and make the users achieve them, and 5 minutes to fill the questionnaire, for a total of 15 minutes for each experiment.

The questionnaire was realized based on the current literature (Ives and Olson, 1984; Zviran and Erlich, 2003) and it considers the two main aspects that a task oriented conversational agent should cover, namely the Quality of Service and the Quality of Experience. The questionnaire used a Likert Scale (Graham et al., 2013) (*never, rarely, sometimes, often, always*) to evaluate each of the following questions:

1. The system was efficient in accomplishing the task.
2. The system quickly provided all the information that I needed.
3. The system is easy to use.
4. The system was incoherent when I interacted using a non-standard or unexpected input.
5. The system has a fluent dialogue.
6. The system was flexible to my needs.
7. I am satisfied by my experience.
8. I would recommend the system.
9. The system is charming.
10. I enjoyed the time that I spent using the system.

During the experiments we asked for feedback from the users, and what came out was that there should be more correlation between the tasks that we asked to users to accomplish, and the questions of the questionnaire. In addition, it would be interesting to add more questions in order to cover more aspects of the interaction that could be perceived and evaluated by the user. In order to do that, we should even better study the task that we ask the users to accomplish.

4.2 Linguistic Stress Tests

A stress test operates a substitution in a user utterance in order to test the behavior of the system in unconventional situations of interaction. Starting from the user interactions described in Section 4.1, we were able to collect the audio and textual logs of the interactions, which were in turn used to model our stress tests. We have analyzed and studied the logs of the conversations obtained by the users' test of each system. According to the literature (Danieli and Gerbino, 1995; Ruane et al., 2018), we proposed three categories of linguistics tests: spelling substitutions, lexical substitution and syntactic substitutions. In total we defined eleven tests, which were divided as follows:

- *Spelling substitutions*: it aims to test the system behavior when words are misspelled or confused, including cases of wrong speech recognition

  (ST-1) Confused Words (e.g. substitute "there" with "their", or "a fianco" and "affianco").

  (ST-2) Misspelled Words (e.g. substitute "accommodation" with "acommodation").

  (ST-3) Character Replacement (e.g. substitute "don't" with "dont", or "po'" with "po", or "lui dà" changed to "lui da").

  (ST-4) Character Swapping (e.g. substitute "casualmente" with "casuamlente" ,or "therefore" with "therefroe").

- *Lexical substitutions*: it aims to test the system behavior when a word is substituted with a less common word or with a more complex expression, preserving the meaning of the utterance

  (ST-5) Less frequent Synonyms (e.g. substitute "home" with "habitation").

  (ST-6) Synonyms specific to a register of speech or a geographical region (e.g. substitute "buongiorno, vorrei mangiare in un ristorante indiano" with the less formal "c'e un posto indiano").

  (ST-7) Coreference (e.g. substitute "Rome" with "the capital of Italy").

- *Syntactic substitutions*: it aims to test the system behavior when a less common grammatical structure of the utterance is used

(ST-8) Active-Passive Alternation (e.g. substitute “I would like to block the credit card” with the less common “I would like that the credit card is blocked”).

(ST-9) Inverted Order of Nouns and Adjectives (e.g. substitute “un piatto di pasta” with the less common “pasta un piatto”).

(ST-10) Anaphora Resolution (e.g. following the system question “did you say Rome or Milan?” substitute “Milan” with the less natural “the second”).

(ST-11) Verbal Modifier Inversion (e.g. substitute “I would like to buy a ticket to Milan for tomorrow” with the less used “I would like to buy a ticket for tomorrow to Milan”).

In order to apply as many tests as possible, among the ones listed above, even when it was not possible to use the log to obtain a suitable test, we created *ad hoc* tests. Before the application of the ad hoc tests, we checked whether the system was able to achieve the task in a non-stressful situation and, afterwards, we applied the stressful condition to our input. As far as spoken dialogue systems are concerned, it was not possible to apply the character replacement test, since it is a condition that can be tested only in a textual context.

5 Qualitative Analysis and Discussion

After having assessed the systems behaviour, we have set up an evaluation report for each of the system. The report includes the following sections: Evaluation summary, Detailed evaluation: questionnaire, and Detailed evaluation: stress test. We now briefly present the content of the three sections, using the CH1 system as example.

Evaluation summary. Here we give an high level statement about each of the three aspects reported in Table 1 (task completion, Effectiveness of the dialogue, user satisfaction). The statement reports the performance obtained on both the questionnaire and the stress tests. For instance, the following is the statement received by CH1 as far as the Effectiveness of the dialogue is concerned:

Effectiveness of the dialogue:

- Questions 3, 4, 7 and 10 of the questionnaire: average CH1 score is 2.03/4

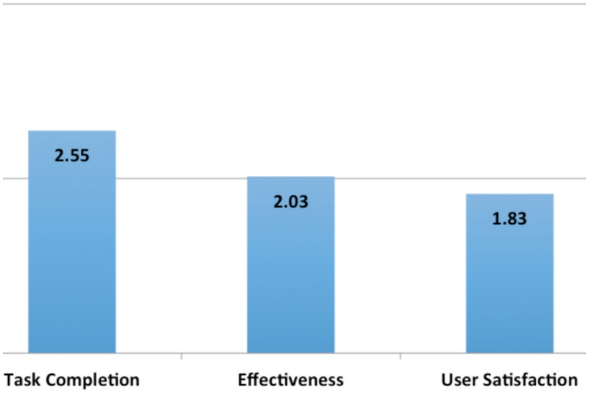

Figure 1: Example report of IDIAL questionnaire.

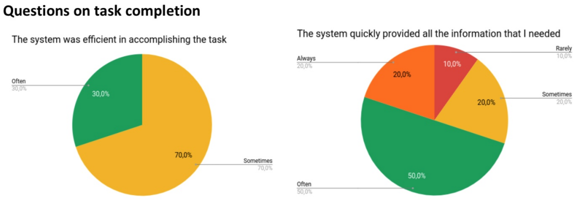

Figure 2: Example report of IDIAL questions on task completion.

- Linguistic stress tests: average CH1 score on the three groups (spelling substitutions, lexical substitution and syntactic substitutions) is 0.59/1.

Detailed evaluation: questionnaire. This section of the IDIAL evaluation reports about the questionnaire on the three aspects. Figure 1 shows the synthetic view on the three aspects of the system, while Figure 2 provides the diagrams reported for the two task completion questions.

Detailed evaluation: stress tests. This section of the IDIAL evaluation reports about the application of the linguistic stress tests. Figure 3 shows

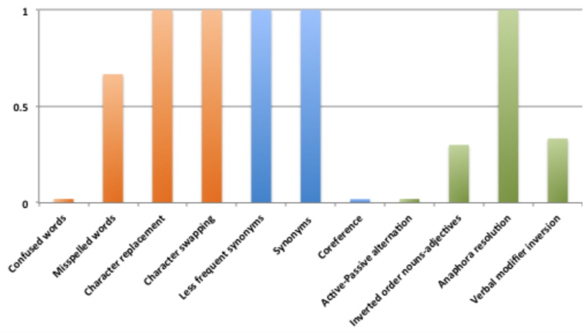

Figure 3: Example report of IDIAL linguistic stress tests.

the CH-1 performance on the eleven stress tests (average success rate (scale 0-1) on the eleven stress tests applied).

6 Post-Evaluation Questionnaire

One of the main aim of the IDIAL task at Evalita is the development of a scalable and domain independent methodology for assessing the performance of conversational agents. In this perspective, we were interested to know how the IDIAL evaluation protocol is perceived by the developers of the conversational agent participating in the task. As a first step in this direction, we submitted a post-evaluation questionnaire to the participants. The questionnaire comprised five questions, as follows:

- How do you judge the evaluation methodology used for IDIAL?

    a) the user experience questionnaire?

    b) the linguistic-oriented stress tests?

- How do you explain the successes or failure of the examined linguistic features?
- Are there aspects of your system which should be better considered in the IDIAL protocol?
- Which evaluation system do you normally use to test the functionality of your system? Which is its reference literature?
- Would you use the IDIAL evaluation protocol as the official metrics for evaluating your systems? Why? Eventually, after what kind of adjustments?

The double nature of the IDIAL protocol, which not only tests the user satisfaction but also proves the effectiveness of the system interactions in unconventional linguistic contexts of use, was generally perceived as a good choice for testing conversational agents. As a matter of fact, stress tests are judged to be a good starting point to improve the quality of linguistic performances for each system. Moreover, this kind of framework is seen to be particularly adequate to compare different systems accomplishing similar tasks.

On the other hand, participants stated that the results returned to them, although being graphically clear and understandable, were not fully satisfying, mainly as far as the variety of the tested interaction situations is concerned. This limitation was particularly relevant for systems showing many interaction capabilities and tasks (i.e. intents): for such systems poor performance on the tested tasks might not be representative of the overall behaviour of the system.

As a second feedback that we received, results in the evaluation report should be enriched with textual explanations, in order to better describe the reasons of failure.

Among other suggestions provided by the participants in the post-evaluation, we mention the need of including a comparison between the expected time of task completion and the actual time spent by users in accomplishing the requested service, the need of extending the number of intents to test, the ability of distinguishing the system ability to understand different entities within the same utterance, and the need of testing different system modules separately (i.e. ASR, TTS, SLU).

7 Conclusion

IDIAL (Evaluation of Italian DIALogue systems) is the first shared task aiming at assessing interactive characteristics of conversational agents for the Italian language. IDIAL considers a dialogue system as a "black box" (i.e., evaluation can not access internal components of the system), and measures the system performance on three dimensions: task completion, effectiveness of the dialogue and user satisfaction. The IDIAL evaluation protocol includes both a questionnaire with subjective user judgments, and a set of linguistic stress tests applied to interactions. The long term goal is the development of a scalable and domain independent methodology for assessing the performance of conversational agents.

Being the evaluated systems different with respect to the task they have been designed to address, the output of the IDIAL evaluation can not be a ranking. Conversely, for each system, we provide an evaluation report with a set of qualitative assessments based on a detailed and coherent set of interactive characteristics of the system. The method is flexible, since both the questions of the questionnaire and the stress tests can be adapted and personalized respecting the general principles of the methodology.

As for future improvements, there are few main aspects that need attention. First, the method should better test the variety of intents covered by the system. The selection we made in our evaluation is not fully representative of the interaction situations of a complex system. Second, the relation between the time of task completion and the actual time spent by users in accomplishing the requested service is not considered in the current protocol. Finally, it would be interesting to deeply test the IDIAL protocol with dialogue systems with differ interaction modalities.

References

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39.

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, Sara Falcone, and Bernardo Magnini. 2018. A methodology for evaluating interaction strategies of task-oriented conversational agents. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 24–32.

Mikko Hartikainen, Esa-Pekka Salonen, and Markku Turunen. 2004. Subjective evaluation of spoken dialogue systems using ser vqual method. In *Eighth International Conference on Spoken Language Processing*.

Blake Ives and Margrethe H Olson. 1984. User involvement and mis success: A review of research. *Management science*, 30(5):586–603.

Bernardo Magnini, Vevake Balaraman, Mauro Dragoni, Marco Guerini, Simone Magnolini, and Valerio Piccioni. 2018. Ch1: A conversational system to calculate carbohydrates in a meal. In *Proceedings of the 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018)*.

Sebastian Möller and Nigel G Ward. 2008. A framework for model-based evaluation of spoken dialog systems. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 182–189. Association for Computational Linguistics.

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.

Elayne Ruane, Théo Faure, Ross Smith, Dan Bean, Julie Carson-Berndsen, and Anthony Ventresque. 2018. Botest: a framework to test the quality of conversational agents using divergent input examples. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, page 64. ACM.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *arXiv preprint arXiv:1701.03079*.

Moshe Zviran and Zippy Erlich. 2003. Measuring is user satisfaction: review and implications. *Communications of the Association for Information Systems*, 12(1):5.