Conversation Analysis, Repair Sequences and Human Computer Interaction - A theoretical framework and an empirical proposal of action

Conversation Analysis, Repair Sequences and Human Computer Interaction

A theoretical framework and an empirical proposal of action

Francesca Alloatti [1, 2], **Luigi Di Caro** [2], **Alessio Bosca** [1]
[1]CELI - Language Technology
[1]Department of Computer Science, University of Turin
Turin, Italy
francesca.alloatti@celi.it

Abstract

Significant increase in neural networks accuracy may not indicate that the system is able to understand language phenomena that we, as humans, consider to be basic features of our communication capability. This work explores the hypothesis that in order to achieve a better exchange in the human-machine dialogue, it is not strictly necessary to improve the understanding ability of the agent; it is important though to teach the machine to recognize a deviation in the normal course of an interaction and to repair it. Conversational Analysis (CA), a discipline that pertains to Sociolinguistics, has studied in detail the ways humans communicate with one another and how they manage to constantly detect any issue in the conversation. CA has compiled a series of repair strategies that people employ to correct those issues. The proposal of action is to transfer some of these findings in the computational domain. This work is relevant on two sides: on one hand, it allows for a more meaningful evaluation of a dialogue agent. On the other hand, it provides a practical plan of action to act upon the mistakes in a conversation.

1 Introduction

In recent years, neural dialogue models - the core of some of the most famous dialog systems - have seen a constant improvement. One of the most used metric to evaluate dialogue system is BLEU (Papineni et al. 2002). However, BLEU was created to evaluate machine translation and it was proved to be ineffective in the domain of conversational agents (Deriu et al. 2019).

Some work has already been done towards establishing other evaluation protocols (Mehri and Eskenazi 2020; Finch and Choi 2020; Takanobu et al. 2020). Cervone and Riccardi (2020) apply findings from applied Linguistics - Discourse Analysis (DA) (Hall and Fine 1977) and the Speech Act Theory (SAT) (Searle 1969) - in order to provide a more complete evaluation. DA and SAT, however, do not offer the right framework to *act* upon the incoherence, or errors, that could be encountered in a dialogue.

In this context, Conversation analysis (CA) proves itself useful for dialogue systems evaluation because of its findings about *repair sequences*. Through the use of CA, this work proposes a methodology to detect breakdowns in conversation, as well as a way to repair them by taking into account the users' expertise and mental model.

2 Background

In this section we present the theoretical background which constitutes the foundations of our work.

2.1 Conversation Analysis

Conversation analysis (CA) relies on the theory that meaning is constructed through the alternation between speakers, and that this alternation, known as *turn-taking*, is regulated by rules provided by the social environment (Sacks, Schegloff, and Jefferson 1974). Norman and Thomas (1991) suggest that Conversation analysis is able to highlight two important characteristics of human exchange: the fact that it is orderly, and that it allows reciprocal intelligibility between speakers.

From the work of CA emerges that a conversation is usually composed by different pairs, called *adjacency pairs*. The pairs do not need to be strictly subsequent one to the other, as they may be split over a sequence of turns. Examples of Adjacency pairs include Questions-Answers, Apology-Acceptance, Complaint-Excuse, etc. (Schegloff and Sacks 1973). Some sequences of turn can accomplish a specific social functions for the speakers; it is the case of *repair strategies*. Repair is the act of correcting any mistake that may happen in a conversation that prevent meaning to be shared fluidly by the two parties.

2.2 Problems and Repairs

In CA, an utterance can be marked as ambiguous not because it pertains to a specific class of ambiguous statements, or because it is ambiguous per se, but because it is interpreted as such by the speakers in the conversation. Once one of the participant observes a discrepancy in the normal course of the conversation, a repair sequence can be initiated. The goal of repairing is to apply the principles of CA in order to bring the conversation back into a structure that allows to build shared meaning.

Most of CA researchers believe that there is no clear connection between the source of error and the type of re-

pair (Schegloff, Jefferson, and Sacks 1977). According to them, the recurrent repair patterns are not bound to their causes: repair work is structured when it occurs, but predicting the points in a conversation where it is most likely to emerge is impossible. However, Higashinaka et al. (2017; 2015) proved that error detection is possible in chat-oriented systems, and even proposed a taxonomy of errors based on the utterance, response or context level in which they appeared. Since task-oriented systems usually follow a predetermined path, it is even more feasible to detect recurring errors in conversations by analyzing existing data: if similar conversation turns always trigger a repair sequence from the second speaker, then problematic patterns become evident. Some work has already been done in this direction (Aberdeen and Ferro 2003; Green et al. 2006) although none employed findings from CA to identify errors. Since it is possible to detect repair sequences automatically in a conversation, by employing CA methodology it is also possible to evaluate the success of that dialogue. Secondly, if problematic spots are known, specific repair strategies can be studied in order to correct the issues and create a more intelligent agent.

3 Derailment Detection and Repair

In this section, we present our modular approach to detect and repair errors in conversations.

3.1 Breakdowns Detection

In this paper, we elaborated a new tagset inspired from the MALTUS one (Popescu-Belis 2004). MALTUS was deemed to be the closer one to our purposes, since it already provided different tags to differentiate various kinds of reaction to errors in a conversation. Other works also used similar kind of tags but they were not as structured as the MALTUS one (Batliner et al. 2003), or contained less tags (Lopes et al. 2015; Cevik, Weng, and Lee 2008), or relied on additional information such as prosodic speech features to detect the breakdowns (Litman, Swerts, and Hirschberg 2006). Not all the tags in the MALTUS tagset were used, since some were not applicable in our context. Table 1 states the tags that were deemed useful. The aim is to apply sequential implicativeness in order to recognize each user input that initiates a repair strategy, since it is only from the second speaker reaction that we can notice a derailment in the conversation.

From an empirical perspective, each answer given by the system should await different kinds of answers: either it's a compliant one, or and understandable one, and then the system can provide a new answer according to its internal reasoner; but if it's a non-compliant input (identified by one of the MALTUS tags), or an attempt from the user to correct the system, then the system must understand this attempt and react appropriately. The system should always be prepared to receive a non-compliant input time and therefore be ready to initiate a repair sequence.

3.2 Repair Initiation

It is not necessary to provide a different answer, or script a different agent behaviour each time the user initiates a repair sequence. A solution would be to provide the answer

Table 1: A selection of useful tags from the MALTUS tagset.

Tag	Description
S	statement
Q	question
RP	positive answer
RN	negative answer
RIC	restated information with correction
RIR	restated information with repetition
DO	command or other performative (includes: command, commitment, suggestion, open option, explicit performative)
PO	the utterance is related to politeness (sympathy, apology, downplayer)
RU	the utterance is related to rudeness (swearing, cruel irony)

that specific user most probably needs, according to his *expertise level* and *mental model*. Within the field of psychology, the definition of *expertise* has encompassed a range of ideas, such as the "extent and organization of knowledge and special reasoning processes to development and intelligence" (Feltovich and Hoffman 1997). The *Mental model*, on the other hand, denotes what the users believe about the system itself, its functions and its internal mechanisms. For instance, expert users may be willing to try different formulations in order to extract information from a system: they know that conversational agents do not have the same understanding capabilities than humans. In this case, it is probably worth to try to provide some clearance over why a certain answer was provided by the system, giving insights into the mechanisms of the agent to the users. Less expert users, on the other hand, may just feel that the system does not understand them for unknown reasons. In this case, it might be sensible to provide a different repair strategy (e.g. give guidance through a simple help button, or a menu of available options). If an intent is identified as a repair request, the system then proceeds to understand what kind of repair request is, according to the tags in the MALTUS tagset.

Each and every one of the tags is paired with an appropriate response. However, it does not mean any answer could be triggered at any time, or that all the repair requests are always valid. Each repair "slot" can be activated only if it fits the expertise assessment and mental model of that user. The act of repairing only makes sense if it actually helps the other speaker: a conversational agent that asks a clarification question to a user with a low expertise and mental model score, would only confuses the user more. The various slots should be activated accordingly to that user assessment, which is calculated in a multilevel fashion integrating expertise and mental model. It is also possible that some sequences are completely blocked by that user's assessment: for instance, in the case of particularly low score, it might not worth trying to repair the interaction.

3.3 User Assessment

Expertise and *Mental model* features assess the users from multiple perspectives, integrating information about their

technical competence as well as their general knowledge about dialogue systems.

User Expertise can assimilate to the notion of User Modeling. There are certain attributes that compose the user status and help categorizing the users in two main groups (Hassel and Hagen 2006): *novices* (non-expert users) and *experts*. The attributes must take into account different dimensions. Namely, a technical knowledge and a general competence.

- **Technical knowledge**. A novice user will employ general concepts, while an expert one will use more technical and appropriate words. For instance, a novice user could often say "I don't know" when the system asks questions related to technical aspects of that service. Moreover, the main topics of a task-oriented dialogue system can be arranged hierarchically from the simplest ones to the more technical and complex, creating various levels of expertise. Each user input can then be classified at a certain expertise level by looking at the words it contains or how the technical concepts were formulated (Ribeiro et al. 2016; Jokinen and Kanto 2004).
- **General competence**. Novice users will anthropomorphize the system (Luger and Sellen 2016). They will point out to human features or personality traits that do not really exist, e.g.by using female or male pronouns while referring to the agent, or by characterizing it with adjectives such as "kind", "rude", "stupid", etc. They may also call out the system by its name, if the agent was given one. The use of a vocative can cover multiple functions, such as establishing a social relationship between the speaker and the addressee (Leech 1999). Novice or unskilled users usually make typos while writing or their utterances may present some agrammatical features characterized as syntactical mistakes.

For the Mental model, the users can be located along a spectrum: on one end, the model of the system as it really is; it could be defined as the Mental model of those who developed the agent, and therefore we call these users *Developers*. On the other end, the model of those who have an opposite perspective, who may be called *Primitives*. It is worth mentioning that these two categories were designed with the specific goal of modeling users of a task-oriented written dialogue system. Other studies were able to incorporate emotions drafted from speech signals, meaning the system was provided with a vocal interface (Callejas, Griol, and Lopez-Cozar 2011). Some attributes that can help locate a user on the spectrum relate to the users' compliance, the use of deictics, the production of out-of-context input and greetings:

- **Compliance.** The first feature is whether users are compliant to the agent's requests. Compliant users, i.e. Developers, use the same words or phrasings that were employed by the system, they press a button when proposed to, or they write a number when they are asked to (Candello, Vasconcelos, and Pinhanez 2017). The compliance can be computed by looking at the overlapping between the agent's phrases or words and the users' ones.
- **Deictics.** Users who heavily employ deictics (e.g. "that day", "the other thing I said", "in my area", "my wife said", etc.) are to be considered Primitives, since they do not understand that even advanced systems have trouble understanding contextual references.
- **Out-of-context.** Primitives may ask out-of-context question, or make unreasonable requests. For instance, they may ask to be called back by a system that does not provide that service, or ask questions that are outside the scope of that agent (e.g. commercial enquiries to an agent that is supposed to provide technical assistance). The detection of unreasonable requests can be done by looking for some peculiar expressions (such as "I want to be called back"), or by detecting out-of-scope utterances.
- **Greetings.** Primitive users often employ formulaic expressions, such as farewell or welcoming sentences while writing to a conversational agent. That is, they apply the same politeness strategies that they use while writing to actual human beings. In this case this is not only useless, but dangerous: adding more noise to the input can only confuse the understanding component of the dialogue agent.

It is worth noting that although the whole system could be applied to all sort of task-oriented agents, user's profiling shall always depend on the specific context and requirements of the agent's domain, specifically for the technical expertise evaluation. Therefore, user evaluation experiments will always have to deal with the exact content of a dataset. The challenge is to create a system that is able to compute errors on the fly, and select the appropriate repair strategy by taking into account the information (Expertise and Mental model) about the user. We believe that this strategy could significantly improve the intelligence of the system, without the need to intervene on the neural dialogue model.

4 Conclusion

This work presents a theoretical framework to evaluate and improve task-oriented dialogue agents. The framework exploits a sociolinguistic discipline, Conversation analysis, and its findings about human interaction. The application of CA structures can help evaluate dialogue systems in a more precise and complete way. The classification of repair strategies (their occurrences in the dialogue, as well as their typology) can show how many times the agent is failing, in what parts of the dialogue, and why. Once the system is able to detect repair strategies on the fly, it can then act and provide a meaningful contextual answer to that request. The notion of meaningful answer strictly depends on the user who is making that request. To that end, users should be classified taking into account their expertise (technical and general) and their mental model of the conversational agent.

The authors are currently applying this framework by classifying breakdowns in a proprietary dataset, where they obtained promising results. Future work will surely include the extension of the task to open datasets, as well as the integration of the hypothesis about the users' expertise and mental model.

References

Aberdeen, J., and Ferro, L. 2003. Dialogue patterns and misunderstandings. In *Proceedings of the Error Handling in Spoken Dialogue Systems workshop*.

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J.; and Nöth, E. 2003. How to find trouble in communication. *Speech Commun.* 40(1–2):117–143.

Callejas, Z.; Griol, D.; and Lopez-Cozar, R. 2011. Predicting user mental states in spoken dialogue systems. *EURASIP J. Adv. Signal Process* 6.

Candello, H.; Vasconcelos, M.; and Pinhanez, C. 2017. Evaluating the conversation flow and content quality of a multi-bot conversational system.

Cervone, A., and Riccardi, G. 2020. Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 162–174. 1st virtual meeting: Association for Computational Linguistics.

Cevik, M.; Weng, F.; and Lee, C.-H. 2008. Detection of repetitions in spontaneous speech in dialogue sessions. In *INTERSPEECH-2008*, 471–474.

Deriu, J.; Rodrigo, Á.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; and Cieliebak, M. 2019. Survey on evaluation methods for dialogue systems. *CoRR* abs/1905.04071.

Feltovich, P. J., and Hoffman, R. R. 1997. *Expertise in context*. AAAI Press Menlo Park, CA.

Finch, S. E., and Choi, J. D. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 236–245. 1st virtual meeting: Association for Computational Linguistics.

Green, A.; Eklundh, K. S.; Wrede, B.; and Li, S. 2006. Integrating miscommunication analysis in natural language interface design for a service robot. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4678–4683.

Hall, W. S., and Fine, J. 1977. J. mch. sinclair and r. m. coulthard. towards an analysis of discourse. london: Oxford university press, 1975. *Language in Society* 6(2):296–299.

Hassel, L., and Hagen, E. 2006. Adaptation of an automotive dialogue system to users' expertise and evaluation of the system. *Language Resources and Evaluation* 40(1):67–85.

Higashinaka, R.; Funakoshi, K.; Araki, M.; Tsukahara, H.; Kobayashi, Y.; and Mizukami, M. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 87–95. Prague, Czech Republic: Association for Computational Linguistics.

Higashinaka, R.; Funakoshi, K.; Inaba, M.; Tsunomori, Y.; Takahashi, T.; and Kaji, N. 2017. Overview of dialogue breakdown detection challenge 3. In *Proceedings of the DSTC6 - Dialog System Technology Challenges*.

Jokinen, K., and Kanto, K. 2004. User expertise modeling and adaptivity in a speech-based E-mail system. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 87–94.

Leech, G. 1999. *The distribution and function of vocatives in American and British English conversation.* Rodopi. 107–118.

Litman, D.; Swerts, M.; and Hirschberg, J. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics* 32(3):417–438.

Lopes, J.; Salvi, G.; Skantze, G.; Abad, A.; Gustafson, J.; Batista, F.; Meena, R.; and Trancoso, I. 2015. Detecting repetitions in spoken dialogue systems using phonetic distances. In *INTERSPEECH-2015*, 1805–1809.

Luger, E., and Sellen, A. 2016. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 5286–5297. New York, NY, USA: Association for Computing Machinery.

Mehri, S., and Eskenazi, M. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–235. 1st virtual meeting: Association for Computational Linguistics.

Norman, M., and Thomas, P. 1991. Informing HCI design through conversation analysis. *International Journal of Man-Machine Studies* 35:235–250.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Popescu-Belis, A. 2004. Dialogue act tagsets for meeting understanding: an abstraction based on the damsl, switchboard and icsi-mr tagsets. Technical report.

Ribeiro, E.; Batista, F.; Trancoso, I.; Lopes, J.; Ribeiro, R.; and de Matos, D. M. 2016. Assessing user expertise in spoken dialog system interactions. *Lecture Notes in Computer Science* 245–254.

Sacks, H.; Schegloff, E.; and Jefferson, G. 1974. A simple systematic for the organisation of turn taking in conversation. *Language* 50:696–735.

Schegloff, E., and Sacks, H. 1973. Opening up closings. *Semiotica* 8:289–327.

Schegloff, E.; Jefferson, G.; and Sacks, H. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53:361–382.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Takanobu, R.; Zhu, Q.; Li, J.; Peng, B.; Gao, J.; and Huang, M. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 297–310. 1st virtual meeting: Association for Computational Linguistics.