Annotating Errors and Emotions in Human-Chatbot Interactions in Italian

Annotating Errors and Emotions in Human-Chatbot Interactions in Italian

Manuela Sanguinetti[♡◇] **Alessandro Mazzei**[♡] **Viviana Patti**[♡]
Marco Scalerandi[♡] **Dario Mana**[♣] **Rossana Simeoni**[♣]
[♡]Dipartimento di Informatica, Università degli Studi di Torino, Italy
[◇]Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Italy
[♣]TIM, Torino, Italy
[♡]{first.last}@unito.it, [◇]{first.last}@unica.it
[♣]{first.last}@telecomitalia.it

Abstract

This paper describes a novel annotation scheme specifically designed for a customer-service context where written interactions take place between a given user and the chatbot of an Italian telecommunication company. More specifically, the scheme aims to detect and highlight two aspects: the presence of errors in the conversation on both sides (i.e. customer and chatbot) and the "emotional load" of the conversation. This can be inferred from the presence of emotions of some kind (especially negative ones) in the customer messages, and from the possible empathic responses provided by the agent. The dataset annotated according to this scheme is currently used to develop the prototype of a rule-based Natural Language Generation system aimed at improving the chatbot responses and the customer experience overall.

1 Introduction

Conversational agents, or chatbots, have become a widespread technology used by an ever increasing number of companies. Most often they are used to automate tasks such as customer support inquiries, providing a 24/7 livechat service via web-based interfaces or standalone mobile apps and enabling natural language interactions with the users. Chatbot systems typically include Natural Language Understanding (NLU) modules aimed at identifying users' intents and extracting the relevant information; these systems either resort to a set of predefined response templates or use Natural Language Generation (NLG) techniques to reply to the user.

Despite the ever increasing chatbot abilities to recognize user requests and provide appropriate answers, miscommunication is still a common issue in human-chatbot interactions (Sheehan et al., 2020). Such shortcomings go hand-in-hand with the need to strengthen, also and especially in customer support contexts, the ability of chatbots to interact in a human-like fashion, so as to encourage the users to engage more in the conversation and to comply with the agent's requests. An improved interaction quality also passes through the creation of emotion-aware chatbots able to recognize the emotional state of the customer and to respond appropriately, showing some form of empathy (Ma et al., 2020).

The work presented here forms part of a wider project consisting in the development of the prototype of a rule-based *genuine* NLG module (Van Deemter et al., 2005). The aim of this module is to improve the answers provided by the template-based chatbot of an Italian online customer service. Its main feature is that it takes as input a set of annotated dialogues between the chatbot and the customers, and it substitutes the template responses provided by the chatbot system, taking into account also various dimensions in the generation process, such as possible errors in the conversations and the presence of emotions (especially negative ones) in the user messages. The intended purposes in building this module are the enhancement of the chatbot responses on multiple levels. From a strict content viewpoint, a selection procedure inspired by narrative roles in explanation (Biran and McKeown, 2017) has been developed in order to generate more relevant and *ah hoc* responses to individual customers' requests as described in Di Lascio et al. (2020). In addition, an emotional layer has been added in the module design to generate empathic responses, more in line with customers' feelings and expectations.

The development of task-oriented dialogue systems for Italian has raised great interest (especially from private companies) and there are high expectations for a new generation of dialogue systems that can naturally interact and assist humans in real scenarios. However, work on Italian language systems remains limited. Some related task with a focus on evaluation has been recently proposed in the context of EVALITA, the periodic evaluation campaign of NLP and speech tools for Italian (Cutugno et al., 2018). As this highlights, developing synergies between public institutions and private companies is extremely important to improve existing solutions for the Italian language, in particular with regard to effectiveness of the dialogue, user satisfaction and societal consequences and risks of dealing with conversational agents. Our aim with this contribution is also to partially fill this gap, presenting the result of a corpus-based study of a set of real human-chatbot conversations carried out in the context of a virtous example of collaboration between academia and private companies.

In this paper, we describe a multi-layer annotation scheme developed to annotate the aforementioned dialogues in the Italian language. The main novelty of this scheme is that it focuses on two dimensions: conversation errors and emotions. The design of the scheme has been driven by two main objectives: 1) to provide a qualitative analysis of these interactions, using a sample dataset to explore the possible interrelations between errors and emotions in conversation, thus providing the company with the opportunity to establish adequate error handling strategies, 2) to feed the NLG module with this additional information so as to improve the newly-generated responses. The distinguishing feature of the proposed scheme is that, in order to make sense of the complexity of the tagset relating to the possible errors, the error types have been clustered in a coarse-grained taxonomy whose classes are partially inspired by the popular Gricean maxims upon which the cooperative principle is built (Grice, 1989).

After outlining past related work, we hereby introduce the dataset used to apply the scheme. We then provide an overview of the scheme itself with some basic statistics and examples. Finally, we provide a qualitative analysis of the annotated data, discussing the main issues arisen during the annotation process.

2 Related Work

Error analysis and classification in both spoken and written human-machine interactions has been a matter of study for decades and its goal is mainly to define an improved set of error recovery strategies or to predict eventual conversation breakdowns. In Skantze (2005), a scheme was developed to establish the degree of comprehension of a task-oriented spoken dialogue system, along with the effects on the user experience. User utterances were thus annotated according to how well they were understood by the system, using the following set of tags: *Non-understanding, Misunderstanding, Partial Understanding, Full Understanding*. More recently, an improved taxonomy of errors was proposed in Higashinaka et al. (2018), based on the revision and comparison of two previous taxonomies (Higashinaka et al., 2015a; Higashinaka et al., 2015b) built according to a theory-driven, top-down approach the former, and with a data-driven, bottom-up approach the latter. In this revision process, both taxonomies were applied to the datasets used for the shared tasks on automatic breakdown detection (Higashinaka et al., 2016); these datasets consist of dialogues between humans and chat-oriented systems. The schemes were meant to identify systems' errors within conversations. The revision process and eventual comparison of the schemes showed a higher reliability, in terms of inter-annotator agreement results, of the bottom-up taxonomy. While the above-mentioned contributions focused on system's errors, a past work by Bernsen et al. (1996) aimed specifically at studying user errors, i.e., those cases where the user fails to comply with the normative model of the dialogue (e.g., not following the system's explicit instructions or deliberately refusing to be cooperative). Finally, in Möller et al. (2007), both parties are taken into account while designing an error annotation scheme within the context of a task-based spoken interaction. In fact, a set of errors on different levels (goal, task, command, concept) was defined for user utterances, but also a recognition-level error was introduced to capture system's mistyping and automatic speech-recognition errors. Interestingly, each user-system exchange was also annotated with a label aimed at capturing the consequences of such errors, which may be *stagnation*, *repetition* (as a special case of stagnation), *regression*, or rather *partial* or *complete progress*.

Parallel to error analysis is the annotation of human emotions in dialogues. While a large number

of contributions on open-domain, human-human interactions can be found in literature (see Poria et al. (2019) for an overview), relatively fewer works are available on the annotation of human-machine dialogues, and more specifically in task-based domains, such as customer service. In addition, the distinguishing purpose of such systems – namely to complete a given task, be it booking a flight of providing assistance of any kind – often makes traditional schemes, such as those based on Ekman o Plutchik theories (Ekman, 1992; Plutchik, 1982), partially inadequate to provide a proper emotion representation within such contexts. As a result, work on this subject mostly focuses on negative emotions that may result from partial or complete failure in completing the given task. Typically, the range of negative emotions include *anger* (Liscombe et al., 2005; Herzig et al., 2016; Schmitt et al., 2012), *frustration* (Ang et al., 2002; Liscombe et al., 2005; Herzig et al., 2016) – even providing fine-grained distinctions between different degrees of emotions (Liscombe et al., 2005; Schmitt et al., 2012) – but also *confusion, disappointment* and *sadness* (Herzig et al., 2016). While such schemes usually apply to user emotions only, recent works also include agent's behavior, proposing annotation labels, as well as computational approaches that aim at identifying empathy on agent's responses (Alam et al., 2018; Herzig et al., 2016).

The annotation scheme described in this paper draws inspiration from the works briefly outlined in this section, and its main contribution lies in the attempt to encompass within a single framework all these different, though in a sense complementary, dimensions.

3 Dataset Description

The scheme has been applied to a corpus of interactions between humans and a task-oriented dialogue system[1]. More specifically, it contains dialogues between customers and a chatbot from the customer care unit of an Italian telecommunication service provider. The dialogues, which take place by means of a text chat, mainly deal with requests for commercial assistance, both on landline and mobile phones.
As briefly introduced in Section 1, the corpus development and its annotation have been designed so as to improve the NLG module of the chatbot, with a particular focus on explanations. Therefore, the dataset was created by selecting, from a sample of dialogues held over 24 hours, a reduced subset that included requests for explanations from customers. The resulting corpus consists of 142 dialogues, including 1540 turns – an average of about 11 turns per dialogue – and an average length of 9 tokens in customer turns and 38 tokens in the agent turns. Similar to Herzig et al. (2016), consecutive messages of the same party (customer or agent), if not interrupted by another message of the other party, were considered as a single turn. This explains the difference in the average turn lengths, as, especially in the case of the chatbot, a single turn may actually consist of more than one message. This difference is also due to the way the agent's responses are currently structured; agents' responses usually include detailed information (for example, on invoice items or available options), while customers' messages are generally more concise. In some cases, the latter are basic yes/no answers, or digits (1,2,...) corresponding to the options provided by the agent in its previous message, as in the example below:

Chatbot *Mi risultano 2 linee a te intestate: 1.* `phoneNumber` *2.* `phoneNumber`*.*
A quale linea devo fare riferimento?
Scegli una delle linee indicate, selezionando la posizione nell'elenco (1,2,3...).
(I've found two numbers on your name: 1.`phoneNumber` 2. `phoneNumber`
Which one should I refer to?
Choose one by selecting the position in the list (1, 2, 3, ...))

Customer 1

The types of requests for explanations collected in this corpus basically reflect the different kinds of problems typically encountered with a telecom service provider, such as undue or unfamiliar charges in the bill or in the phone credit (these cases represent about 52% of the overall number of requests in this dataset). An issue emerged from the data in the corpus is the occurrence of breakdowns in conversation, i.e., whenever the interaction is interrupted because one or both parties give up the conversation without completing the task (Martinovsky and Traum, 2003). In fact, out of the 142 conversations in this

[1] Due to privacy reasons, the corpus cannot yet be publicly released, but we are currently working on its anonymization.

collection, 53 have been interrupted by the customer, who did not get to a proper answer or solution to the problem posed. As for the remaining conversations, only in 10 cases the conversation had a positive outcome (the chatbot provided a satisfactory explanation or proper help), while the other 79 ended with the handover to a human agent, which was often explicitly asked for by the customer.

Besides providing a rough indicator of the interaction quality, this prompted further analysis in order to explore, under different perspectives, the causes of such breakdowns and the dynamics behind these interactions. The analysis led to the development of a richer and fine-grained annotation scheme aiming to capture these aspects, as described in the next section.

4 Annotation Scheme

This section provides an overview of the tagset adopted for the annotation of both dimensions included in the scheme, i.e. errors and emotions. Detailed guidelines with examples can also be found in the following document (in both Italian and English): `https://cutt.ly/cdMcnyM`

4.1 Errors

The main motivation behind the design of this annotation scheme was to provide a detailed account of the possible errors that could be encountered in human-chatbot interactions, especially within a customer service domain. By error, in this context we mean any event that might have a negative impact on the flow of the interaction, and more in general on its quality, potentially resulting in conversation breakdowns. Such events usually represent a deviation from an expectation, both by the customer and the agent, or more simply a deviation from standard linguistic norms, which may also result in misinterpretations of any kind.

The tagset was conceived so as to include a wide range of phenomena on different levels and, unlike other taxonomies of errors specifically defined for just one party in the interaction (either user or system, as described in Section 2), it includes error categories that may apply to either only customer's or only chatbot's messages, or to messages from both. While certain errors can be attributed to only one party, others can be considered as transverse phenomena, and we are interested in providing a comprehensive view capturing all of them. This motivates the high variety of tags defined in the scheme. Despite this variety, however, we clustered the error types in a coarse-grained taxonomy whose classes are partially inspired by the popular Gricean maxims upon which the cooperative principle is built (Grice, 1989), more specifically to the maxim of Quantity, Relation and Manner[2]. We then define a more generic class of errors. The terminology used to define the error tags is partially borrowed from Higashinaka et al. (2018) and Bernsen et al. (1996), whose contributions focused only on system errors (the former) and on user errors (the latter).

Below we describe the general characteristics of the error classes. Table 1 reports the complete list of the error tags according to their corresponding class and it indicates whether they are specific to messages from one party (customer or agent) or if they can be applied to both.

Quantity This class includes all those messages that violate the maxim of Quantity, according to which contributions should be as informative as required, but not more informative than required. Therefore, with this class we define those messages in which insufficient information is provided to obtain a relevant response ("Lack of Information"), or, conversely, in which the text is unnecessarily verbose and redundant ("Excess of Information"). The error tags in this class are meant to be used for both parties.

Relation This class includes a large number of error types that signal the violation of the maxim of Relation (i.e. "*Be relevant*") in various ways. The messages annotated with these tags do not have a relation to the previous message to various degrees: they might state something that only implicitly

[2]The first maxim, that of Quality ("*Try to make your contribution one that is true*", (Grice, 1989), p.26) has not been considered here, as we assume that it is in the interest of both parties to be truthful and always provide the correct information.

	Customer	Chatbot
Quantity	Lack of information	
	Excess of information	
Relation	Indirect response	
	Ignoring question/feedback	
	Repetition	
	Straight wrong response	
		Topic change
	Answering with question	
Manner	Ill-formed	
	Indirect question	
	Non-understandable	
		Grammatical error
Generic	Non-cooperativity	
	Other	

Table 1: List of the error tags according to their corresponding class. Errors are displayed on the left, right or center column depending on whether they pertain messages from customer, chatbot or both.

answers the previous question ("Indirect response") or just provide an answer that seems not to take into account what was stated in the previous message ("Ignoring question/feedback"), also re-proposing the same content of past messages ("Repetition")[3], or giving an answer that is completely irrelevant with respect to the previous message ("Straight wrong response"). In case of misunderstandings, the chatbot could propose to switch to another topic ("Topic change"). It is also possible that the customer replies to a request by the chatbot with another question ("Answering with question").

Manner The third class includes messages that violate the maxim of Manner ("*Be perspicuous*") and, in line with Grice's definition, it does not aim to draw attention to what is said but to *how* it is said. Therefore the error tags falling under this class mainly highlight flaws in the message form, rather than in its content. In this case, most of the errors are defined only for customer messages. In fact, the chatbot is a typical template-based dialogue system, and it uses a set of manually created responses, formulated so that they are as clear as possible, unambiguous and grammatically correct (except for some rare cases that we decided to report with the addition of the tag "Grammatical error"). On the customer side, it is more likely to find cases in which the message, although potentially relevant at that stage of the conversation, does not correspond to the form expected by the system (see the "Ill-formed" tag, which mostly applies to cases where the user fails to select the provided options in the expected way), it contains statements that may ambiguously sound like implicit requests ("Indirect question"), or it is nearly or completely obscure ("Non-understandable"). It is worth pointing out that while customers' messages may also contain grammatical errors, the corresponding tag was not used for such cases due to the fact that the user's text, as it appears in the dataset, is usually revised at run-time by a built-in module of the dialogue system that automatically corrects several typos and other basic grammatical errors. The NLU module of the system thus receives as input a potentially revised version of the message that we are not able to see, therefore we cannot determine whether the user's message still contains errors or not.

Generic The last class includes two final tags: one is defined for those messages where the customer deliberately acts in a non-cooperative manner, refusing to provide the system with the information it requests ("Non-cooperativity"); the "Other" tag is finally used for messages presenting errors that do not fit in any of the other categories in this taxonomy.

[3]In this sense, "Repetition" can be considered as a specification of the "Ignoring question/feedback" tag.

	% **Customers**	% **Chatbot**
Quantity	3,43	1,07
Relevance	13,57	14,05
Manner	7,43	0,36
Generic	2,86	0,60
None	72,71	83,93

Table 2: Distribution of customers' and chatbot errors (%) by class over the total number of customers' and chatbot turns.

Customers Errors	**%**
Ignoring question/feedback	20.94
Ill-formed	13.09
Repetition	12.04
Excess of information	11.52
Indirect response	8.90
Non-cooperativity	8.90
Answering with question	8.90
Non-understandable	5.24
Straight wrong response	4.71
Indirect question	3.14
Other	1.57
Lack of information	1.05

Table 3: Distribution of error types among customers errors only.

Chatbot Errors	**%**
Topic change	30.37
Repetition	23.70
Straight wrong response	20.74
Ignoring question/feedback	6.67
Indirect response	5.93
Lack of information	4.44
Other	3.70
Excess of information	2.22
Grammatical error	2.22

Table 4: Distribution of error types among chatbot errors only.

Table 2 shows the error distribution in the customers' and chatbot messages, organized by main classes, while Tables 3 and 4 summarize the distribution of the specific error types computed over the total number of customers' and chatbot errors respectively (thus ignoring the largest portion of turns where no error occurs).

4.2 Emotions and Empathy

The annotation of the emotional dimension is meant to identify for each turn in the conversation both the customer's emotion and the possible signals of empathic response in the chatbot message.

Customer's emotions Due to the context where the interactions take place, i.e. a customer service, we selected a limited set of the possible range of emotions that can be typically applied to a corpus of conversations (such as in IEMOCAP (Busso et al., 2008), DailyDialog (Li et al., 2017) or EmoContext (Chatterjee et al., 2019), to name a few). We thus followed a similar scheme as the one adopted in Liscombe et al. (2005), including a generic "Positive" emotion that simply expresses a positive attitude towards the conversational agent and a "Neutral" tag to indicate whether no particular emotion is perceived in the message. We then introduced two negative emotions, *frustration* and *anger*, providing for both a fine-grained distinction with the tags "Somewhat/Very frustrated" and "Somewhat/Very angry". Frustration is defined as a "*key negative emotion that roots in disappointment, [...] an irritable distress after a wish collided with an unyielding reality*" (Jeronimus and Laceulle, 2017). In this context, frustration can derive from a sense of disappointment for not receiving the answers sought or a solution to the problem posed, and more generally, from an experience with the chatbot that is unsatisfactory and does not correspond to the expectations. The user can experience frustration whenever s/he feels that the interaction with the virtual agent is not leading anywhere. In this sense, the feeling of frustration can

Emotion	%
Somewhat angry	35.35
Somewhat frustrated	32.32
Very frustrated	9.09
Other negative	9.09
Positive	8.08
Very angry	6.06
Empathy	0,95

Table 5: Distribution of emotions detected in customers' turns and of empathic responses by the chatbot (%). The relative frequency of customers' emotions is computed discarding the share of neutral instances.

then be associated to signs of annoyance and dissatisfaction from the customer.

Anger, on the other hand, is intended as a negative emotion that impels aggressive behavior (Novaco, 2017), and it can be triggered, among other factors, by the perception of an unfair treatment. It is especially the idea of being treated unfairly that motivated the definition of the corresponding tag in our scheme. We thus labeled as "anger" any covert or overt hostility (e.g., when the customer threatens to end the use of the service) namely derived from this perception.

The two degrees of anger and frustration are distinguished on the basis of lexical and pragmatic cues (e.g., in the use of specific words or interjections) identified in the customer's message or other cues inferred from the context of the exchange. We finally added the "Other negative" tag for negative emotions that are neither anger nor frustration. The choice of including only two main negative emotions (excluding others used in previous works, as mentioned in Section 2) is motivated by an exploratory analysis of the data in the pre-annotation phase in which anger and frustration were found to be the two prevailing emotions. To avoid further sparseness on an already reduced dataset, we focused on these two emotions and added a third label to conflate any other negative emotions.

Agent's Empathy The operational definition of empathy is borrowed from the work by Alam et al. (2018), who define empathy as "*a situation where an agent anticipates or views solutions and clarifications, based on the understanding of a customer's problem or issue, that can relieve or prevent the customer's unpleasant feelings*" (p.40). We found this definition suitable for this work as it applies to a fairly similar context, that is a collection of dyadic Italian task-oriented dialogues[4].

The inclusion of this tag was motivated by our interest in studying the agent's behavior and its ability to properly recognize and respond to customers' signs of dissatisfaction. The very low distribution of empathic responses as reported from the dataset, shows that there is much room for improvement in this respect, and the NLG module that is currently under development within this project should address this issue as well. Table 5 shows the distribution of emotions and empathy in the dataset. In the next section we describe the scheme usage while commenting on a few annotated examples.

4.3 Annotation Examples

Below, we report two dialogue excerpts[5] in which some of the tags in the scheme were used. The turn number in the actual conversation is written in subscript on the left.

[4]In Alam et al. (2018), however, the data at issue consists of human-human interactions.

[5]For privacy reasons, customers' personal information such as invoice or phone number are replaced with a placeholder.

Example 1

Chatbot$_6$ *Ecco il dettaglio delle voci di spesa presenti nella fattura* `InvoiceNumber` *del mese di gennaio per la linea* `PhoneNumber`: [...] *Vuoi avere il dettaglio di una specifica voce di spesa presente in fattura?*
(Here is the detail of the `InvoiceNumber` of January for the `PhoneNumber`: [...] Do you want the detail of a specific item in the invoice?)

Customer$_7$ *Son rimasta senza internet 16 giorni a dicembre..non devo pagare questi giorni*
(I was left without internet for 16 days..I don't have to pay for those days)

Error: Ignoring question/feedback -Emotion: Somewhat angry

Chatbot$_8$ *Vuoi cambiare argomento?*
(Do you want to change topic?)

Error: Topic change

Customer$_9$ *Voglio una fattura giusta per il servizio che ho avuto*
(I want a fair invoice for the service I had)

Emotion: Somewhat angry

Chatbot$_{10}$ *Scusami, non ho capito bene la tua richiesta... Puoi ripeterla in modo più semplice?*
(I'm sorry, I haven't understood your question very well... Could you please rephrase it in a simpler way?)

Customer$_{11}$ *Voglio restituiti i soldi del servizio che non ho avuto per 16 gg*
(I want the money back for the service I haven't had for 16 days)

Error: Repetition - Emotion: Somewhat angry

Chatbot$_{12}$ *Verifico quali addebiti ci sono stati sul tuo* `PhoneNumber`
(I check which charges have been made on your `PhoneNumber`)

Error: Straight wrong response

Breakdown

Example 2

Chatbot$_3$ *Scusami, non ho capito bene la tua richiesta... Puoi ripeterla in modo più semplice?*
(I'm sorry, I haven't understood your question very well... Could you please rephrase it in a simpler way?)

Customer$_4$ *No non hai capito niente*
(No, you've got it all wrong)

Error: Non-cooperativity - Emotion: Somewhat frustrated

Chatbot$_5$ *Mi spiace non aver compreso correttamente. Ho la possibilità di farti chattare con un operatore umano più esperto. Vuoi?*
(I'm sorry I haven't understood that correctly. I can let you chat with an expert human representative. Do you want me to?)

Empathy

In Example 1, the customer contacts the service to ask for explanations regarding a charge in the bill. The chatbot reports the details of the invoice and it offers to provide additional information for each amount, but the customer responds in an irrelevant way, ignoring the agent's question and insisting on the problem, showing signs of anger, due to the feeling of being treated unfairly. This "sense of injustice" also persists in the customer's subsequent messages, with the message in turn 11 repeating what already said in turn 7. When the customer explicitly requests for a refund and the agent texts back with an incorrect response related to a charge check, the customer eventually leaves the conversation.

Example 2 shows a case of customer non-cooperativity, with the customer refusing to repeat the question (not reported in the excerpt) in a way that is more understandable to the system. In this context,

we interpreted the customer's non-cooperative attitude as a sign of frustration for an interaction that did not bring the expected results. The chatbot response in turn 5, instead, was evaluated as sufficiently empathic, for it offers an apology and proactively suggests the possibility to switch to a human agent (who is expected to provide a higher-quality assistance).

4.4 Qualitative Analysis

Although the dataset used for the annotation experiments is quite small to draw ultimate conclusions, we were still able to get valuable insights on the main dynamics underlying these customer-chatbot interactions. First, the error distributions reported in Tables 2–4 shows that both customers and chatbot mostly tend to produce messages that violate the maxim of Relevance, i.e. they produce messages that are irrelevant to the context of the conversation. In the case of customers, the most frequent errors within this class are those in which the customer does not take into account the previous message posted by the chatbot ("Ignoring question/feedback"), and the one in which the same content appears in multiple messages ("Repetition"). On the chatbot side, the most representative error of this class is the one where the chatbot asks the user to switch to another topic ("Topic change"). This is a typical expedient used by the system when it does not properly recognize the user's intent, and it is relevant in this context as, in turn, in 53% of cases this is the result of an error in the preceding message by the user (mostly "Ignoring question/feedback" and "Indirect response" errors).

With a further analysis, we also observed the co-occurrence of customers' errors and negative emotions in an attempt to provide an explanation of the errors produced by the customers and, conversely, to find an empirical basis of the emotion definitions we eventually elaborated in our guidelines. We found that the type of customers' errors that mostly co-occur with negative emotions are "Non-cooperativity" (83.3% of cases), "Excess of information" (50%) and "Repetition" (43.5%). Due to the higher sparsity, we discarded the other types. For such errors, frustration is the prevailing negative emotion in "Non-cooperativity" and "Repetition", while anger is the one that mostly occurs in the remaining one. Unlike the relation between "Non-cooperativity" and "Repetition" on one side, and frustration on the other, the association between anger and "Excess of information" seemed less intuitive and required a manual inspection of concrete examples from the dataset. What emerged is that the longer messages that typically occur in this error type are often due to the fact that the customer is experiencing several problems and on different levels (e.g., undue charges, poor service quality, inefficiencies, etc.). This contributes to increase customers' perception of unfair treatment and anger.

5 Agreement Results and Discussion

The first round of the annotation process was carried out by two independent annotators, following the guidelines. The Inter-Annotator Agreement (IAA) statistics, measured using the Cohen's k coefficient (Cohen, 1960), were then computed after this step, and, as expected, they reported from moderate to low results. More specifically, we obtained a k=0.46 for error categories, k=0.43 for the agent's empathy and k=0.28 for customers' emotions. The resulting annotation was then reviewed to solve the cases of disagreement, which were discussed by an extended team including two other partners of the project.

The goal of this annotation work was to have the broader coverage possible of the problems encountered in a conversational context such as that of a customer service chat. Hence, the need to create a large and varied tagset. However, this variety has been at the expense of annotation consistency, as the IAA results show. Despite the presence of detailed guidelines, the subjectivity of the individual annotators proved to be crucial, and not only - as one would have expected - with respect to the choices regarding emotions and empathy, but also in the error annotation. In the latter, in fact, the highest disagreement was reported where, for the same turn, one annotator identified an error while the other did not, thus showing different perceptions of the dialogue state and its deviation from a regular path. With a similar trend, most of the disagreement in the annotation of emotions was detected when an annotator identified an emotion of some kind in the customer's message, whereas the other one perceived the message as completely neutral. The second main cause of disagreement for this dimension was the different perception of anger and frustration between the two annotators. Despite an agreed-upon definition of

both emotions, in practice, different interpretations were often given to the users' messages and to the conversational contexts in this respect. By contrast, the fine-grained distinction between somewhat/very angry or frustrated did not cause substantial disagreement between the annotators and the conflation of the two labels into a single one, both for anger and frustration, did not significantly improve IAA results, while conflating Frustration and Anger into a single label increased the agreement rate to k=0.40. These findings brought to a revision of the definition of both emotions in the guidelines.
It is worth pointing out that when designing the scheme for this dimension, we took into account the fact that customers' interventions could be affected by the awareness of the medium used. For example, customers may consider that politeness is not required or that their messages should be as simple as possible for the machine to understand them and, accordingly, their interventions may result quite straightforward and could be wrongly believed to carry some kind of anger. Instead, what we actually observed in the sample set is that most often – maybe due to little familiarity with such tools – customers are not fully aware of this. They keep posing questions or answering to the chatbot's questions as if they were interacting with a real person. We hypothesize that frustration and anger may arise even due to this factor. An in-depth study will be carried out to verify if there is enough evidence to confirm it.

Agreement results on empathy are also moderate, even more so if compared to the ones obtained in Alam et al. (2018), who reported a Cohen's k=0.74, and whose definition of empathy has been used in this work. We attribute such divergence to the differences in the type of data collected, as Alam et al. (2018) focus on human-human spoken interactions, while in this work we deal with conversations between humans and a chatbot system with a relatively limited set of predefined responses. Whether the same response, in different conversational contexts, could be considered as sufficiently empathic was thus left at the annotator's discretion. All these issues prove once again the complexity of this task, especially considering that it deals with interactions on a text chat, rather than via spoken conversations, in which other factors such as acoustic and prosodic features can help discriminate. This is also the reason why the reviewing phase to solve these cases was extended to two other group members who had not participated in the first step, so as to reach a final version as widely shared as possible.

6 Conclusions

In this paper, we described the main features of an annotation scheme designed to label a sample set of interactions between customers and the chatbot of an Italian telecom customer service unit. The ultimate goal of this scheme and its applications on the sample data is to develop a rule-based NLG module to be integrated within the template-based chatbot system and intended to substitute template responses with more context-dependent responses, with an eye in particular on the improvement of explanations. With this purpose in mind, a set of dialogues was selected, where the customer requests for an explanation of some kind. The specific characteristics of these conversational contexts - errors of various nature from both parties, potential customer dissatisfaction as perceived within the dialogues, etc. - as well as the expected output of the NLG module under development motivated the way the scheme was devised.

The annotation experiments carried out to validate the scheme highlighted the significant role played by subjectivity of individual annotators, which is reflected by a moderate IAA, but also the need to further analyze the inherent meaning of some of the labels that proved to be more controversial (as, for example, the ones introduced to define anger and frustration in customers' messages), and to revise their usage accordingly. The annotation layer devoted to make the presence of negative emotions in the customer's message explicit was designed with the aim of improving the chatbot's awareness of the user's emotional state. Indeed, this information is essential as a basis for generating an emotion-aware and empathic response, which is the subject of future work. On this line, we also aim at investigating how to manage the cases of frustrated or angry users who insult the chatbot, in order to avoid escalation or responses where the chatbot shows himself unaware of the insult (Cercas Curry and Rieser, 2019).

Acknowledgements

The work of Alessandro Mazzei, Manuela Sanguinetti and Viviana Patti has been partially funded by TIM s.p.a. (*Studi e Ricerche su Sistemi Conversazionali Intelligenti*, CENF_CT_RIC_19_01).

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language*, 50:40 – 61.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*. ISCA.

Niels Ole Bernsen, Laila Dybkjær, and Hans Dybkjær. 1996. User errors in spoken human-machine dialogue. In *Proceedings of the ECAI '96 Workshop on Dialogue Processing in Spoken Language Systems*, ECAI '96, Berlin, Heidelberg. Springer-Verlag.

Or Biran and Kathleen McKeown. 2017. Human-centric justification of machine learning predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1461–1467. International Joint Conferences on Artificial Intelligence Organization.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden, September. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Francesco Cutugno, Maria Di Maro, Sara Falcone, Marco Guerini, Bernardo Magnini, and Antonio Origlia. 2018. Overview of the EVALITA 2018 evaluation of italian dialogue systems (IDIAL) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mirko Di Lascio, Manuela Sanguinetti, Luca Anselma, Dario Mana, Alessandro Mazzei, Rossana Simeoni, and Viviana Patti. 2020. Natural language generation in dialogue systems for customer care. In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-It 2020)*. CEUR-ws.org.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachussets.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73, Los Angeles, September. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. 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*, pages 87–95, Prague, Czech Republic, September. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Lisbon, Portugal. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2018. Improving taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*.

Bertus F. Jeronimus and Odilia M. Laceulle. 2017. Frustration. In Virgil Zeigler-Hill and Todd K. Shackelford, editors, *Encyclopedia of Personality and Individual Differences*, pages 1–5. Springer International Publishing, Cham.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Jackson Liscombe, Giuseppe Riccardi, and Dilek Hakkani-Tür. 2005. Using context to improve emotion detection in spoken dialog systems. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, pages 1845–1848. ISCA.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50 – 70.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Dialogue Systems*.

Sebastian Möller, Klaus-Peter Engelbrecht, and Antti Oulasvirta. 2007. Analysis of communication failures for spoken dialogue systems. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 134–137. ISCA.

Raymond W. Novaco. 2017. Anger. In Virgil Zeigler-Hill and Todd K. Shackelford, editors, *Encyclopedia of Personality and Individual Differences*, pages 1–5. Springer International Publishing, Cham.

R. Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21:529 – 553.

S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3369–3373. European Language Resources Association (ELRA).

Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115:14 – 24.

Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325 – 341.

Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24, March.