Content Selection for Explanation Requests in Customer-Care Domain

Luca Anselma[♡] **Mirko Di Lascio**[♡] **Dario Mana**[♣]
Alessandro Mazzei[♡] **Manuela Sanguinetti**[♡♢]
[♡]Dipartimento di Informatica, Università degli Studi di Torino, Italy [♣]TIM, Torino, Italy
[♢]Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Italy
[♡]{first.last}@unito.it, [♢]{first.last}@unica.it
[♣]{first.last}@telecomitalia.it

Abstract

This paper describes a content selection module for the generation of explanations in a dialogue system designed for customer care domain. First we describe the construction of a corpus of dialogues containing explanation requests from customers to a virtual agent of a telco, and second we study and formalize the importance of a specific information content for the generated message. In particular, we adapt the notions of *importance* and *relevance* (Biran and McKeown, 2017) in the case of schematic knowledge bases.

1 Introduction

Customer care is one of the application domains where Dialogue Systems (DSs) represent an emerging technology used by many big companies to satisfy customer requests (MITTR, 2018). Customer care dialogues can have a specific linguistic characterization (Oraby et al., 2019), and often the customer preferences lean toward short dialogues (Demberg et al., 2011). Moreover, in the customer care domain the users' requests often regard some form of *explanation* about their past transactions with the company. To provide explanations, commercial DSs often provide long lists of data entries extracted from databases containing company-customer relationship data. Therefore, there is the necessity to give some form of *priority* to data entries to present just – or to give more prominence to – the information that is most relevant for the user (Demberg et al., 2011).

Most commercial DSs follow the classical cascade architecture $NLUnderstanding \leftrightarrow DialogueManager \leftrightarrow NLGeneration$ (McTear et al., 2016). This architecture relies, as a working hypothesis, on the assumption that most of necessary information is provided by the user utterance. However, this assumption is sometimes false or only partially true. For instance, in the sentence "*Scusami ma vorrei sapere come mai mi vengono fatti certi addebiti?*" ("Excuse me, I'd like to know why I'm charged certain fees?"), even a very advanced NLU module can produce only vague information about the user's request to the dialogue manager. Indeed, to provide an appropriate response, the dialogue manager might need to ask for additional clarification or, in alternative, to access some contextual information to obviate the lack of linguistic information. In the case of customer care, this contextual information can be found as schematic knowledge bases arising from databases. As a result, when linguistic information is scarce (or *absent* in the case of ungrammatical/incomprehensible input) retrieving and giving priority to contextual information in DSs is essentially a problem of *content selection* (Reiter and Dale, 2000). Therefore, as a working hypothesis, in this paper we consider negligible the linguistic input given by the user. However, also when the linguistic input is comprehensible, a good balance between the information carried by the linguistic input and by the specific domain context is a key goal for the dialogue manager.

The idea to use NLG techniques for explaining rationales inside data is a topic that is drawing growing attention (Reiter, 2019). One of the few papers providing a quantitative evaluation of explanations was produced by Biran and McKeown (2017). In this work the authors proposed a model for quantifying the relevance of a feature for a specific class of machine learning algorithms, i.e. linear classifiers. The authors introduced two notions, *importance* and *effect*, to evaluate the relevance of a feature in the general classification model and for a specific classification instance respectively. The basic idea was to determine the narrative role of a feature based on the combination of its importance and its effect; for example, a feature may

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 5–10,
Dublin, Ireland, 18 December 2020.

have the narrative role of *exceptional evidence* in the case of low importance and high effect. In this way, the authors have been able to communicate the key data elements into core messages for an explanation (*justification* in their terminology).

In this paper, we present some initial results of an ongoing study on the design of a generation module of a DS in the domain of telco customer care. We focus our study on customers' requests of *explanations* (Reiter, 2019). The study presented here, in fact, is part of a wider project that aims to improve the answers provided by a virtual agent of an online customer service, by creating a NLG system that could also take into account various dimensions in the generation process, such as possible errors in the conversations (see e.g. Bernsen et al. (1996), Martinovsky and Traum (2003), Higashinaka et al. (2015)) and the presence of emotions (especially negative ones) in the user messages. At this stage of the project, we use the model presented in Biran and McKeown (2017) to give relevance to the *content units* in the knowledge about the customer. In particular, we adapt the definition of the narrative roles for *importance* and *effect* to the case of a knowledge base consisting of database entries.

This paper provides two main specific contributions: (1) the analysis of a corpus consisting of real dialogues containing explanation requests (Section 2), (2) the proposal of a content-selection procedure based on narrative roles in explanation when the DS contextual data is a schematic knowledge base arising from a database (Section 3). In the final Section of the paper we discuss these contributions in relation to our ongoing work.

2 Building a corpus of explanation requests

This study builds upon the analysis of a corpus of dialogues between customers and a virtual agent for customer service developed by an Italian telecommunications company (Sanguinetti et al., 2020). The dialogues, which take place by means of a textual chat, mainly deal with requests for commercial assistance, both on landline and mobile phones. For the purpose of this study, the corpus created was extracted by selecting, from a sample of dialogues held over 24 hours, a reduced subset that included requests for explanations from customers. The selection criteria were conceived so as to include all the dialogues where at least one message from the user contained a clearly stated

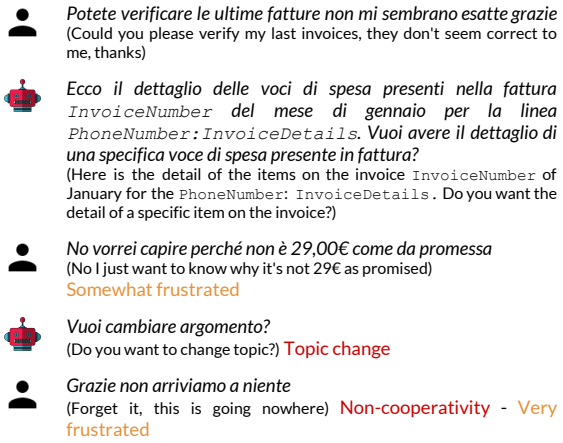

Figure 1: An annotated dialogue with additional annotation layers: errors (red) and emotions (orange).

request for explanation. A simple string search method was thus carried out to filter such kind of dialogues, using the following strings: *sapere/capire perché*[1] ("know/understand why") and *come mai* ("why/how come"). The resulting corpus consists of 142 dialogues, with an average of 11 turns per dialogue, and an average length of 9 tokens in customer messages and 38 tokens in the bot messages. Such difference in the message length is due to the way the assistant's responses are currently structured, in that they usually include detailed information e.g. on invoice items or options available, while, on the other hand, customer's messages are most often quite concise. Also, the relatively high number of turns per dialogue might be explained with the high occurrence in the corpus of repeated or rephrased messages, both by the virtual agent and the user, due to recurring misunderstandings on both sides. The corpus underwent an annotation process that involved multiple, complementary, dimensions, such as errors in conversation and emotions (see Figure 1 for an example).

The explanation request and its sub-types have been included as one of such dimensions and we mainly focused our attention on these in this phase of the study. The types of requests for explanations in this collection reflect the different kinds of problems typically encountered with a telephone operator. Based on a preliminary analysis of the corpus, we distinguished 5 main types of requests plus a generic category that includes a variety of cases that is more heterogeneous and not classifiable according to the main types. Hence, we identified requests for explanations or clarifications

[1] Variants as *perchè, xk, xkè* have been used too.

regarding the following topics: (1) charges in the invoice or in the phone credit (about 52% of cases), (2) timing and methods of receipt of the invoice (10.5%), (3) unpaid invoice reminders erroneously received (10.5%), (4) currently active promotions (8%), (5) payment methods (5%). The remaining cases (14%) were included in the more generic "Other" class. Starting from this analysis we thus defined a reduced set of possible *scenarios*, i.e. prototypical situations that can be found in the dialogues and grouped together according to similar characteristics. For illustrative purposes, we describe in this paper the three main scenarios defined for the first request type, i.e the one regarding undue or unclear charges, being by far the most frequent case of request. In Scenario 1 (31% of the occurrences) the customer asks for an explanation about a higher charge with respect to previous ones, also providing specific information on the amount charged; in Scenario 2 (58% of the occurrences), a charge in the account is claimed, but no further information is provided in the user's message. In Scenario 3 (11% of the occurrences) the customer asks for an explanation about a negative balance.

3 Importance and Effect for Content Selection in the Customer Care Domain

We consider three scenarios arising form the corpus analysis (Section 2). Formally, each scenario consists of a set of sequences of transactions, where a transaction is a money transfer operation between a customer and the company (i.e., an amount paid for a certain service). As a result, each transaction sequence represents the different amounts paid along a time period for a specific service (transaction type). To determine the importance and the effect of a transaction sequence, we assume to know all the transactions on the user's account in the last seven months.

It is worth pointing out that the two most important elements in this specific context are money and time. Therefore, we want to formalize the intuition that the *importance* (in Biran and McKeown's terminology) of a telco service can be associated with the amount of money that the user usually spends for such service, while its *effect* can be associated with the amount of money that the user spent for the service in the last month.

We thus define the *importance of a transaction sequence* as the mean of the normalized values of the transactions in the past six months. Moreover, we define the *effect of a transaction sequence* as the normalized value of the transactions in the current month. Normalization is carried out by dividing the amount of the transactions by the maximum amount that the user has paid for that transaction. An important point in the Biran and McKeown (2017) model is the procedure for transforming importance/effect numeric values in the discrete {low, high} values. In accordance to the original model, we determine the smallest subset H of transaction sequences such that the sum of their importance/effect values is greater than the 75% of the total importance/effect. When such a smallest subset is not unique, we consider the union of all the smallest subsets. Note that the value of 75% has been set in order to adhere to the original model, that has been proposed in (Biran and McKeown, 2017) without a specific motivation. However, we consider this limit as a tunable value that should be empirically validated on the specific domain.

In the following discussions we analyse the scenarios for three common requests of explanation. We separately analyse these scenarios but not that they are not mutually exclusive. It is worth noting that a complete NLG architecture could account their possible coexistence by using some form of syntactic or semantic aggregation.

Finally, note that in the discussion on these scenarios we are completely neglecting both the linguistic information arising from the dialogue (the user's question) as well as any kind of information on the customer. In other words, we are inferring the customer's explanation request as a content selection task only, without taking user utterances and user model into account. As a matter of fact, there are some cases where the user searches for a complete information about its transactions: for instance, the user wants to review all the transactions of the last months. In this case, the linguistic input should trigger the dialogue manager and the NLG system to provide information on transactions with *normal* evidence after the information on exceptional evidence. In contrast, there are situations such that the user wants to have only a short summary on its transactions and in this case the NLG system should only provide information on transactions with exceptional evidence. In future research, we plan to study how to merge the linguistic, the domain and the user model information.

	M1	M2	M3	M4	M5	M6	M7
S1	10	10	10	10	10	10	10
S2	0	0	0	0	0	2	2

Table 1: The distribution of transactions along the current month (M7) and the previous six months (M1-M6) for Scenario 1.

	M1	M2	M3	M4	M5	M6	M7
S1	9.99	9.99	9.99	9.99	9.99	9.99	9.99
S2	0	0	0	0	2	2	2, 2
S3	0	0	0	0	0	0	1.59

Table 2: The distribution of transactions for Scenario 2.

3.1 Scenario 1

Scenario 1 represents a typical situation of a user requesting for an explanation about a total charge in the current month higher than the ones in the previous months. The interaction between the DS and the user starts with a short message: *Salve vorrei sapere perché mi sono stati presi* 12 € *invece che dieci dall'ultima ricarica (Hi I'd like to know why you got* 12 € *instead of ten since last top-up).*

We assume for this scenario that the user paid for two services[2] (that are transaction sequences, see Table 1). In particular, a transaction of 10€ is present in each month (M1-M7) for S1, while a transaction of 2€ is present only in months $M6$ and $M7$ for S2. By using the data in Table 1, we can calculate the importance and the effect for S1 and S2. The importance of S1 is $(10/10 + 10/10 + 10/10 + 10/10 + 10/10 + 10/10)/6 = 1$, while the importance of S2 is $(0/2 + 0/2 + 0/2 + 0/2 + 0/2 + 2/2)/6 = 0.17$, thus the sum of the importance values is 1.17 and its 75% value is 0.88. The smallest subset H_I such that the sum of the importance values of the transactions is greater than 0.88 is $H_I = \{S1\}$. As a result, S1 has high importance, while S2 has low importance. The effects of S1 and S2 are $10/10 = 1$ and $2/2 = 1$, therefore the sum of the effect values is 2 and its 75% is 1.5. The smallest subset H_E such that the sum of the effect values is greater than 1.5 is $H_E = \{S1, S2\}$, hence S1 and S2 have both high effect. Thus, combining the discrete values of importance and effect, S1 is a normal evidence since it has high importance and high effect, and S2 is an exceptional evidence since it has low importance and high effect. This exceptional evidence captures the intuition that S2 is more informative than S1 in Table 1. As a consequence, S2 will have a central role in the requested explanation.

[2]Note that the trivial solution to return both contents does not solve the problem of assigning them a priority in presentation.

3.2 Scenario 2

Scenario 2 represents a user requesting an explanation about some specific charges (*Scusami ma vorrei sapere come mai mi vengono fatti alcuni addebiti — Sorry but I'd like to know why there are some charges*).

This scenario has three transaction sequences: S1, with an amount of 9.99€ (M1-M7), S2 with an amount of 2€ (M5-M7, appearing twice in M7), and S3 with an amount of 1.59€ (M7) (see Table 2). From this data, we calculate importance and effect for S1, S2 and S3, and their narrative roles as described previously. The importance of S1 is 1, the importance of S2 is 0.33 and the importance of S3 is 0. The sum of the importance values is 1.33 and its 75% is 0.99. The smallest subset H_I such that the sum of the importance values is greater than 0.99 is $H_I = \{S1\}$, so S1 has high importance, while S2 and S3 have low importance. The effect of a transaction sequence is given by the values in the current month: S1 and S3 effect is 1 and S2 effect is 2. The sum of the effect values is 4 and its 75% is 3. The smallest subset H_E such that the sum of the effect is greater than 3 is $H_E = \{S1, S2, S3\}$, hence S1, S2 and S3 have all high effects. As a result, combining the discrete values of importance and effect S1 is a normal evidence and S2 and S3 are both exceptional evidences.

3.3 Scenario 3

Scenario 3 represents a user requesting an explanation about a negative balance (*Buongiorno,vorrei sapere perché ho il credito in negativo, nonostante abbia fatto una ricarica da* 15€ *proprio stamattina — Good morning, I'd like to know why I have a negative balance, despite I made a* 15€ *recharge just this morning*).

This user has three transactions sequences: S1 with an amount of 13€ (M1-M3) and 15€ (M4-M7), S2 with an amount of 0.9€ (four times in M7), and S3 with an amount of 1.99€ (M7) (see Table 3). From these data, we can calculate importance and effect for S1, S2 and S3 and their narrative roles

	M1	M2	M3	M4	M5	M6	M7
S1	13	13	13	15	15	15	15
S2	0	0	0	0	0	0	0.9, 0.9, 0.9, 0.9
S3	0	0	0	0	0	0	1.99

Table 3: The distribution of transactions for Scenario 3.

as previously described. The importance of S1 is 0.94, the importance of S2 and S3 is 0. The sum of the importance values is 0.94 and its 75% value is 0.71. The smallest subset H_I such that the sum of the importance values is greater than 0.71 is $H_I = \{S1\}$, so S1 has high importance, while S2 and S3 have low importance. S1 and S3 effect is 1, while S2 effect is 4. The sum of the effect values is 6 and its 75% value is 4.5. The smallest subset H_E such that the sum of the effect values of the transaction sequences in the subset is greater than 4.5 can be $\{S1, S2\}$ or $\{S2, S3\}$. The subset H_E is the union the two cases, i.e. $H_E = \{S1, S2, S3\}$, hence S1, S2 and S3 have high effect. Thus, S1 is a normal evidence, and S2 and S3 are exceptional evidences.

4 Conclusions and Future Work

This paper reports the first results of an ongoing study on the role of NLG for a DS in the customer care domain. We provided a corpus analysis that shed some light on the customer requests regarding explanations[3]. Moreover, we adapted the model proposed in Biran and McKeown (2017) for narrative roles in explanation for this specific kind of input data. In this way, we designed a content selection procedure accounting for *evidence* of data.

We are working on the inclusion of the content selection procedure described in this paper into a complete NLG architecture for DS. In this linguistically sound NLG architecture, we use a simple rule-based sentence planner (Anselma and Mazzei, 2018) in combination with the Italian version of SimpleNLG (Mazzei et al., 2016) for generating messages that give emphasis and priority to the content elements with high evidence. For instance, in this architecture we can decide to generate final messages that contain only (or mention primarily) contents with exceptional evidence.

As a future work, we are designing a user-based comparative evaluation of the DS exploiting the complete NLG architecture following the schema adopted in (Demberg et al., 2011). The idea is to show both a real dialogue from the corpus and a dialogue obtained with the complete NLG architecture, and to ask users to rate each dialogue and compare them by using a number of Likert-scale questions.

[3]We are currently working on the anonymization of the corpus in order to publicly release it.

Acknowledgment

This research has been partially funded by TIM s.p.a. (Studi e Ricerche su Sistemi Conversazionali Intelligenti, CENF CT RIC 19 01).

References

Luca Anselma and Alessandro Mazzei. 2018. Designing and testing the messages produced by a virtual dietitian. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 244–253.

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*.

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.

Vera Demberg, Andi Winterboer, and Johanna D. Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015. 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.

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*.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK. Association for Computational Linguistics.

Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*, 1st edition. Springer Publishing Company, Incorporated.

MIT Technology Review Insights MITTR. 2018. Humans + bots: Tension and opportunity.

Shereen Oraby, Mansurul Bhuiyan, Pritam Gundecha, Jalal Mahmud, and Rama Akkiraju. 2019. Modeling and computational characterization of twitter customer service conversations. *ACM Trans. Interact. Intell. Syst.*, 9(2–3).

Ehud Reiter. 2019. Natural language generation challenges for explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating errors and emotions in human-chatbot interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159, Barcelona, Spain. Association for Computational Linguistics.