# MULTI-STAGE CLARIFICATION IN CONVERSATIONAL AI: THE CASE OF QUESTION-ANSWERING DIALOGUE SYSTEMS

A PREPRINT

**Hadrien Lautraite**
National Bank Of Canada
600 de la Gauchetière
Montréal, Québec
hadrien.lautraite@bnc.ca

**Nada Naji**
National Bank Of Canada
600 de la Gauchetière
Montréal, Québec
nada.aj.naji@gmail.com

**Louis Marceau**
National Bank Of Canada
600 de la Gauchetière
Montréal, Québec
louis.marceau@bnc.ca

**Marc Queudot**
National Bank Of Canada
600 de la Gauchetière
Montréal, Québec
marc.queudot@bnc.ca

**Eric Charton**
National Bank Of Canada
600 de la Gauchetière
Montréal, Québec
eric.charton@bnc.ca

October 29, 2021

**ABSTRACT**

Clarification resolution plays an important role in various information retrieval tasks such as interactive question answering and conversational search. In such context, the user often formulates their information needs as short and ambiguous queries, some popular search interfaces then prompt the user to confirm her intent (e.g. "Did you mean ... ?") or to rephrase if needed. When it comes to dialogue systems, having fluid user-bot exchanges is key to good user experience. In the absence of such clarification mechanism, one of the following responses is given to the user: 1) A direct answer, which can potentially be non-relevant if the intent was not clear, 2) a generic fallback message informing the user that the retrieval tool is incapable of handling the query. Both scenarios might raise frustration and degrade the user experience. To this end, we propose a multi-stage clarification mechanism for prompting clarification and query selection in the context of a question answering dialogue system. We show that our proposed mechanism improves the overall user experience and outperforms competitive baselines with two datasets, namely the public in-scope out-of-scope dataset and a commercial dataset based on real user logs.

***K*eywords** Dialogue systems · conversational search systems · conversational information seeking · clarification · clarifying questions · mixed-initiative · neural networks

## 1 Introduction

Dialogue systems have been increasingly prevalent in many industries with the rise of virtual assistants such as Apple Siri, Amazon Alexa, and Microsoft Cortana. Such conversational agents can perform a variety of tasks such as, making transactions, booking appointments, or answering users' questions, among others [1].

In the context of question-answering agents, we often talk about *intents*, which represent the various information needs or possible questions that users might have. Intents act as classes within the bot Natural Language Understanding (NLU) model. Such a model attempts at associating an incoming user message to one of the predefined intents learned from training data. Upon detection of an intent by the NLU model, the dialogue system will take a corresponding action, specifically, responding to the user with the answer associated with the detected intent. Under the hood, each intent

is represented with several possible formulations in the training data since users can express themselves in a various ways to convey the same thought. As a concrete example, "forgot my password, what to do?" and "how to recover my pass-code" both relay the same intent but are phrased differently. This added complexity means that the intent classifier has to be generalized enough to handle unseen formulations. Since user queries are often short and ambiguous, the model might assign the wrong intent which yields an incorrect answer being given to the user. To address this issue, bots can have a clarification mechanism which engage the user to confirm or clarify their intent.

In this paper, we focus on question-answering dialogue systems. Such agents can act as an additional communication channel that allows clients to ask a variety of questions and could therefore alleviate the pressure on the corporate call center for customer service and assistance. Additionally, dialogue systems constitute a more convenient tool than having to go through multiple possible answers returned by several searches on a traditional search engine. We propose a multi-stage clarification framework that allows to confirm the user intent before answering if the system's confidence is low and a mechanism to suggest some related formulations in case of the user's confirmation being negative. Evaluation on both click data from real interaction logs and human labeled data demonstrates the high quality of the proposed method, outperforming threshold optimization strategies.

The rest of the paper is arranged as follows: the next section outlines related work. Section 3 describes the datasets we used followed by the experimental setup in Section 4. Afterwards, we present and discuss the results of our work in Section 5. Finally, Section 6 concludes our work and discusses future avenues.

## 2 Related Work

Interpretation issues are often the number one recurrent reason of bad user experience in dialogue systems [2]. Such issues translate as incapability of the dialogue system to understand the user request. In order to alleviate such issues, the dialogue system could trigger a fallback mechanism, that is, by answering that it does not understand the query or does not know the answer. Følstad and Brandtzaeg. [2] reports this behavior as the second most source of user dissatisfaction. The clarification process has been studied in various forms. In 1980, McKeown [3] presents a natural language interface to search in a database. The rule base system generate paraphrases to clarify the user intent in order to generate a database query that answer the user's needs.

Users tend to write short queries that are often ambiguous. This makes it challenging for a search engine or a dialogue system to predict possible intents, only one of which may pertain to the user query at hand[4]. Search engines often use diversification to address this issue, by conveying multiple possible intents. Alternatively, the user is asked a question to *clarify* her information need. This latter approach is essential for what is often referred to as "limited bandwidth" interfaces [5], such as speech-only and small-screen devices[4] yet is also found to be beneficial in web search [6]. Such bidirectional interaction lends itself to dialogue systems.

Braslavski et al. [7] studied the different forms of clarification questions asked by humans on online forums such as the community question answering platform, Stack Exchange. The authors classify those questions in different categories including: requests for more information and questions in the form of "have you tried ..." . Recent advances in the field of deep learning offer new possibilities in conversational AI. Several studies propose neural networks architectures to rank possible responses in an information retrieval system [8], [9], [10]. Yang et al. [8] suggest categorizing user intents in forum discussions in classes, namely, original question, clarification questions, feedback or positive answer. The authors propose a new model named Intent-Aware Ranking with Transformers (IART) based on transformers [11] in order to detect user intents and use it as an attention mechanism for ranking possible answers in a dialogue flow. Their proposed method leverages context when to decide whether to ask the user for further information or to provide an answer based on previous interactions. Zamani et al. [12] developed a transformers-based model for ranking or selecting possible clarification question. Other studies [4], [13], directly tackle the task of generating clarification questions. Rao and Daumé III [13] proposed an adversarial approach to generate clarification question.

Asking too many clarification questions comes with a risk of deteriorating user experience due to overly inquisitive behavior. Sekulić et al. [14] studied user engagement with the clarification pane in search engines in order to determine when and how to prompt users for clarification. Peixeiro et al. [15] address the issue from an optimization perspective in order to determine the ideal threshold to maximize the number of correct direct answers for a maximum number of intents which in turn minimizes the number of unnecessary clarification questions.

We propose a simple yet effective method to provide users with the information they need while keeping a balance of direct answers and request for clarification. Instead of using question generation based on real interactions, which could expose us to data leakage [16], we use canonical formulations from intents with similar keywords as clarification questions. Moreover, our proposed method does not require an additional ranking model to sort all possible clarification reformulations but rather use the confidence score from the initial natural language understanding module in order to

rank the candidates canonical formulations. We show that our method improves effectiveness and allows for more fluid interactions. Its simplicity and the fact that no additional data are needed to train and maintain an supplementary ranking model makes our solution easy to deploy in real industrial context.

## 3 Datasets

We conducted our experiments on two datasets. The **first dataset** is based on logs of real user interactions with our in-house corporate dialogue system. The dialogue system is deployed on our corporate transactional web platform. The dataset contains 8768 conversations collected during the first week of November 2020covering 272 distinct intents. We refer to this dataset as HOUSE. The labels are inferred based on user interactions with the dialogue system. That is, when an intent is recognized by the NLU, the dialogue systems confirms the intent with the user "I understand you want to talk about ...", and if the user clicks "yes" then an association is logged between the query and the intent. Such associations are used as ground-truth labels in our experiments. The dataset is mainly in French.

The **second dataset** is a publicly-available one known as the in-scope and out-of scope dataset designed by Larson et al. [17] to train dialogue systems and evaluate their performance levels on a mix of *in-scope* and *out-of-scope* queries. In-scope queries can be mapped to an intent that is already known by the dialogue system (i.e., appears in the training set). Whereas an out-of-scope query represents a new or unknown concept to the dialogue system. For the purpose of our study, we use only the in-scope portion. The intents cover a variety of topics such as travel, banking, and car maintenance among others. Table 1 presents some of the intents with some corresponding training examples. We refer to this datasett as SCOPE. The SCOPE dataset contains training and testing sets. We use the training set of 150 intents with 100 formulations each to train a dialogue system. As the evaluation of the dialogue system's performance with our clarification pipeline requires manual interactions, we focus our testing on the first 30 intents, considering only the first 10 formulations out of 30. The remaining 20 formulations are used as a validation set in order to fine-tune the fallback threshold of the dialogue system used as benchmark.

| Intent | Examples |
|---|---|
| translate | what expression would i use to say i love you if i were an italian<br>can you tell me how to say 'i do not speak much spanish', in spanish |
| transfer | i need $20000 transferred from my savings to my checking<br>complete a transaction from savings to checking of $20000 |
| travel alert | does ireland have any travel alerts i should be aware of<br>does north korea have any travel alerts i should be aware of |
| PTO request | how do i put in a pto request for the first to the ninth<br>am i allowed to put in a pto request for now to april |
| oil change how | how do i change a car's oil<br>can you find instructions on how to change oil in a car |

Table 1: Examples of intents and training samples from the SCOPE dataset

## 4 Methodology

Our dialogue systems are based on the Rasa Open Source framework. The pipeline consists of the following components: Firstly, a pre-processor which performs several NLP steps such as tokenization and featurization of the queries to obtain sparse representations at both word and character levels. The second component is Rasa's own intent classifier DIET [18] with an NLU model that we trained for 200 epochs.

During data preparation, we created a canonical formulation (one sentence) for each intent. This formulation describes the intent in natural language and is displayed to the user in the clarification pipeline to validate what she meant. For instance, "I understand that you want to talk about opening a new account, is that correct?" is the canonical formulation that is attached to the intent *open new account*.

Our proposed multi-stage clarification pipeline encompasses the following stages:

**Stage 0 - Direct Answer**: in this stage, the dialogue system model *understood* the user intent, that is the confidence level of the predicted intent is above the 75% threshold. A direct affirmative response is given to the user.

**Stage 1 - Confirmation**: the dialogue system enters this stage when it is not sure to have understood, that is, the confidence level of the prediction is less than the threshold. Here, the dialogue system displays the canonical formulation