Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition

Floris den Hengst
Vrije Universiteit Amsterdam
f.den.hengst@vu.nl

Ralf Wolter
ING Group NV
Ralf.Wolter@ing.com

Patrick Altmeyer
TU Delft
P.Altmeyer@tudelft.nl

Arda Kaygan
ING Group NV
Arda.Kaygan@ing.com

arXiv:2403.18973v1 [cs.CL] 27 Mar 2024

Abstract

We present Conformal Intent Classification and Clarification (CICC), a framework for fast and accurate intent classification for task-oriented dialogue systems. The framework turns heuristic uncertainty scores of any intent classifier into a clarification question that is guaranteed to contain the true intent at a pre-defined confidence level. By disambiguating between a small number of likely intents, the user query can be resolved quickly and accurately. Additionally, we propose to augment the framework for out-of-scope detection. In a comparative evaluation using seven intent recognition datasets we find that CICC generates small clarification questions and is capable of out-of-scope detection. CICC can help practitioners and researchers substantially in improving the user experience of dialogue agents with specific clarification questions.

1 Introduction

Intent classification (IC) is a crucial step in the selection of actions and responses in task-oriented dialogue systems. To offer the best possible experience with such systems, IC should accurately map user inputs to a predefined set of intents. A widely known challenge of language in general, and IC specifically, is that user utterances may be incomplete, erroneous, and contain linguistic ambiguities.

Although IC is inherently challenging, a key strength of the conversational setting is that disambiguation or *clarification* questions (CQs) can be posed (Purver et al., 2003; Alfieri et al., 2022). Posing the right CQ at the right time results in a faster resolution of the user query, a more natural conversation, and higher user satisfaction (van Zeelt et al., 2020; Keyvan and Huang, 2022; Siro et al., 2022). CQs have been considered in the context of information retrieval (Zamani et al., 2020) but have received little attention in the context of task-oriented dialogue.

Deciding when to ask a CQ and how to pose it are challenging tasks (DeVault and Stone, 2007; Keyvan and Huang, 2022). First, it is not clear when the system can safely proceed under the assumption that the true intent was correctly identified. Second, it is not clear when the model is too uncertain to formulate a CQ (Cavalin et al., 2020). Finally, it is unclear what the exact information content of the clarification question should be.

We present Conformal Intent Classification and Clarification (CICC), a framework for deciding when to ask a CQ, what its information content should be, and how to formulate it. The framework uses conformal prediction to turn a models' predictive uncertainty into prediction sets that contain the true intent at a predefined confidence level (Shafer and Vovk, 2008; Angelopoulos et al., 2023). The approach is agnostic to the intent classifier, does not require re-training of this model, guarantees that the true intent is in the CQ, allows for rejecting the input as too ambiguous if the model is too uncertain, has interpretable hyperparameters, generates clarification questions that are small and is amenable to the problem of detecting out-of-scope inputs.

In a comparative evaluations with seven data sets and three IC models, we find that CICC outperforms heuristic approaches to predictive uncertainty quantification in all cases. The benefits of CICC are most prominent for ambiguous inputs, which arise naturally in real-world dialogue settings (Zamani et al., 2020; Larson et al., 2019).

2 Related Work

We discuss related work on ambiguity and uncertainty detection within IC and CP with language models.

Clarification Questions Various works acknowledge the problem of handling uncertainty in intent classification and to address it with CQs. Dhole

(2020) proposes a rule-based approach for asking discriminative CQs. The approach is limited to CQs with two intents, lacks a theoretical foundation, and provides no intuitive way of balancing coverage with CQ size. Keyvan and Huang (2022) survey ambiguous queries in the context of conversational search and list sources of ambiguity. They mention that clarification questions should be short, specific, and based on system uncertainty. We propose a principled approach to asking short and specific questions based on uncertainty of any underlying intent classifier for the purposes of task-oriented dialogue.

Alfieri et al. (2022) propose an approach for asking a CQ containing a fixed top-k most likely intents with intent-specific uncertainty thresholds. This approach does not come with any theoretical guarantees and its hyperparameters need to be tuned on an additional data set whereas our approach comes with guarantees on coverage of the true intent and with intuitively interpretable hyperparameters that can be tuned on the same calibration set. We do not compare directly to this method but include top-k selection in our benchmark.

CQs have been studied in other domains, including information retrieval (Zamani et al., 2020), product description improvement (Zhang and Zhu, 2021), and open question-answering (Kuhn et al., 2023). In contrast to the task-specific domain investigated in this work, these domains leave more room for asking generic questions for clarification and do not easily allow for incorporating model uncertainty. Furthermore, the proposed methods require ad hoc tuning of scores based on heuristic metrics of model uncertainty, and do not provide ways to directly balance model uncertainty with CQ size.

Uncertainty and out-of-scope detection The out-of-scope detection task introduced by Larson et al. (2019) is a different task from the task of handling model uncertainty and ambiguous inputs (Cavalin et al., 2020; Yilmaz and Toraman, 2020; Zhan et al., 2021; Zhou et al., 2021). However, predictive uncertainty is often used in addressing the out-of-scope detection task. Although the tasks of handling ambiguous input and detecting out-of-scope input are different, we briefly discuss approaches that leverage model uncertainty for out-of-scope detection here.

Various out-of-scope detection approaches train an intent classifier and tune a decision boundary based on a measure of the classifier's confidence (Shu et al., 2017; Lin and Xu, 2019; Yan et al., 2020; Hendrycks et al., 2020). Samples for which the predictive uncertainty of the model lies on one side of the boundary are classified as out-of-scope. These approaches use the models' heuristic uncertainty to decide whether an input is out-of-sample whereas we first turn the models' heuristic uncertainty into a prediction with statistical guarantees and then use this prediction to decide when and how to formulate a clarification question. We additionally propose an adaptation of the CICC framework for out-of-scope detection.

Conformal Prediction on NLP tasks Conformal Prediction has been used in several NLP tasks, including sentiment classification by Maltoudoglou et al. (2020), named-entity recognition by Fisch et al. (2022) and paraphrase detection by Giovannotti and Gammerman (2021). However, the application to intent classification, task-oriented dialogue and the combination with CQs presented here is novel to our knowledge.

3 Methodology

We address the problem of asking CQs in task oriented dialogue systems in the following way. We take a user utterance and a pre-trained intent classifier, and then return an appropriate response based on the predictive uncertainty of the model. Algorithm 1 lists these steps, and an example input is presented in Figure 1. In this section we describe and detail the components of CICC. We start by providing a background on conformal prediction.

3.1 Conformal Prediction

Conformal Prediction is a framework for creating statistically rigorous prediction sets from a heuristic measure of predictive uncertainty of a classifier (Shafer and Vovk, 2008; Angelopoulos et al., 2023). We here focus on split conformal prediction as it does not require any retraining of the underlying model, and refer to it simply as conformal prediction from here on out.

For a classification task with classes $\mathcal{Y} : \{1, \ldots, K\}$, a test input $X_t \in \mathcal{X}$ with label $Y_t \in \mathcal{Y}$, and a user-defined error level $\alpha \in [0, 1)$, CP returns a set $\mathcal{C}(X_t) \subseteq \mathcal{Y}$ for which the following holds (Vovk et al., 1999) even when using a finite amount of samples:

$$\mathbb{P}\left(Y_t \in \mathcal{C}(X_t)\right) \geq 1 - \alpha \tag{1}$$

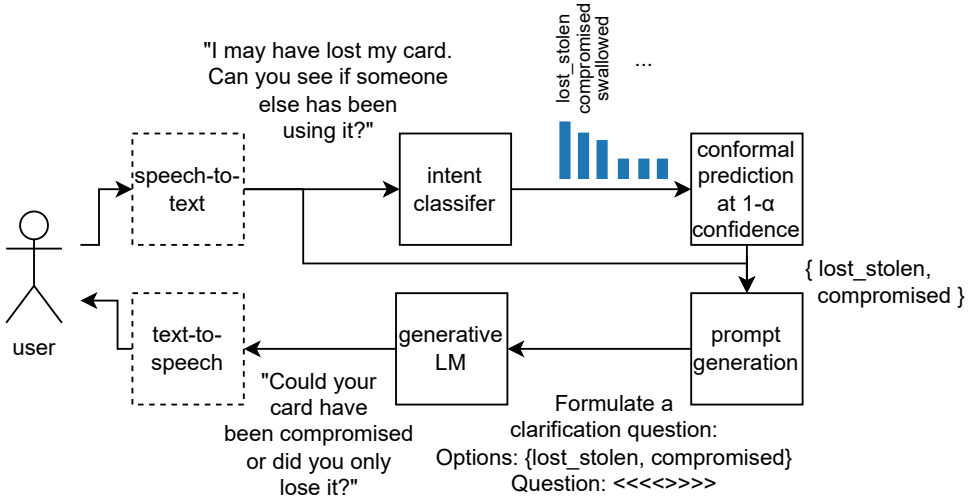

Figure 1: The conformal intent classification and clarification interaction loop.

If e.g. $\alpha = 0.01$ the set $\mathcal{C}(X_t)$ is therefore *guaranteed* to contain the true Y_t in 99% of test inputs.

Conformal prediction uses a heuristic measure of uncertainty of a pretrained model and a modestly sized calibration set to generate prediction sets. Formally, we assume a held-out calibration set $D : \{(X_i, Y_i)\}$ of size n, a pre-trained classifier $\hat{f} : \mathcal{X} \to \mathbb{R}^K$, and a nonconformity function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that returns heuristic uncertainty scores where larger values express higher uncertainty. An example of a nonconformity function for a neural network classifier is one minus the softmax outputs of the true class:

$$s(X_i) := 1 - \hat{f}(X_i)_{Y_i}. \tag{2}$$

This score is high when the softmax score of the true class is low, i.e., when the model is badly wrong.

The nonconformity function s is evaluated on D to generate a set of nonconformity scores $\mathcal{S} = \{s(X_i, Y_i)\}$. Next, the quantile $\hat{q}$ of the empirical distribution of $\mathcal{S}$ is determined so that the desired coverage ratio $(1 - \alpha)$ is achieved. This can be done by choosing $\hat{q} = \lceil (n+1)(1-\alpha) \rceil / n$[1] where $\lceil \cdot \rceil$ denotes the ceiling function. Then, for a given test input X_t, all classes $y \in \mathcal{Y}$ with high enough confidence are included in a prediction set $\mathcal{C}(X_t)$:

$$\mathcal{C}(X_t) := \{y : s(X_t, y) \leq \hat{q}\}. \tag{3}$$

This simple procedure guarantees that (1) holds i.e. that the true Y_t is in the set at the specified confidence $1 - \alpha$. Note the lack of retraining or ensembling of classifiers, that the procedure requires little compute and that D can be relatively small as long as it contains a fair number of examples for all classes and is exchangeable[2] with the test data (Papadopoulos et al., 2002).

There are various implementations of conformal prediction with different nonconformity functions and performance characteristics. The most simple approach is known as *marginal* conformal prediction and it uses the nonconformity function in (2). Marginal conformal prediction owes its names from adhering to the guarantee (1) marginalized over $\mathcal{X}$ and $\mathcal{Y}$, i.e. it satisfies the coverage requirement (1) on average, rather than e.g. for a particular input X_t. Marginal CP can be implemented following the steps described previously: (i) compute nonconformity scores S using (2), (ii) obtain $\hat{q}$ as described previously, and (iii) construct a prediction set using (3) at test time. A benefit of this approach is that it generates prediction sets with the smallest possible prediction set size on average. A limitation is that its prediction set sizes may not reflect hardness of the input (Sadinle et al., 2019).

Alternatively, one can ensure conditional adherence to (1) with so-called conditional or adaptive conformal predictors. A benefit of conditional approaches is that higher model uncertainty results in larger prediction sets. However, a downside is that these sets are expected to be larger on average than those obtained with a marginal approach.Romano et al. (2020) introduce a conditional CP approach that consists of broadly the same steps as marginal CP but with a different nonconformity function s and a different prediction set construction. First we define a permutation $\pi(X)$ of $\{1 \ldots K\}$ that sorts

[1] this is essentially the $\hat{q}$ quantile with a minor adjustment

[2] distributed identically but not necessarily independently