Direct Answer Threshold Optimization in Dialogue Systems

Marco Peixeiro, Nada Naji, Eric Charton *
Banque Nationale du Canada (BNC), Montreal, QC, Canada

Abstract

The presence of dialogue systems is rising in a wide array of industries. While complex, human-like conversational flow and turn-taking have been the focus of recent research advances, we make the point that direct answers are equally important in human-bot interactions. This is especially true in an information seeking task where prompt, correct answers with minimal back-and-forth are desirable. We define a direct answer as a response given to a user query without requiring further clarifications from the user. To determine whether a direct answer is to be given or not, a threshold is applied to the to the confidence level of the predicted intent; in the case where the confidence is higher than the threshold, the user receives a direct answer. This threshold is often set intuitively or on the basis of a few observations, usually between 50% - 75%. In this paper, we propose a method to estimate this threshold based on the intent classification confidence level combined with several intent volumetrics. The goal of our method is to maximize the number of correct direct responses for as many intents as possible in order to minimize user frustration from unnecessary requests for clarifications. Moreover, our method is applicable in the earlier stages of a dialogue system when real interaction logs are scarce. We show that our method improves the accuracy of directly answered queries by 3 to 14% while maximizing the number of accurately answered intents on two dialogue system datasets of 32 and 152 intents.

Keywords: dialogue systems, conversational information seeking, question answering, direct answer threshold, human-chatbot interaction

1. Introduction

Dialogue systems are becoming increasingly present in many industries, allowing the automation of various tasks such as information seeking, and booking appointments, among many other tasks [1]. With a growing offering of dialogue systems by many corporations, consumers' expectations are rising when it comes to conversational AI interactions. One key element in providing the best user experience is the capacity of the dialogue system to provide the user with an accurate, direct answer. A direct answer interaction is defined as a query or natural language question that receives an answer without requiring further information or clarifications from the user [2]. The flipping-point between providing a direct answers and requesting further query clarifications is usually set by a threshold applied to the confidence level of the intent prediction. This threshold is often decided to be *not too high but not too low*, often based on intuition or a few observations rather than statistical grounds. Setting the threshold too low will cause the bot[1] to respond hastily with potentially non-relevant answers. Setting the threshold too high might raise inquisitive behaviour as the bot constantly requires further clarifications. Both scenarios could cause frustration and subsequently user dissatisfaction as they lose trust in the bot and its reliability [3–5]. We propose a systematic approach to determine the optimal response threshold for a dialogue system. We view this as an optimization problem, wherein a trade-off exists between the accuracy of predicting the right intent and the number of intents in a dialogue

[1]Our work considers a customer support question answering scenario. We use "bot", in this context, synonymously to a dialogue system for simplicity.

*marco.peixeiro@bnc.ca *nada.naji@bnc.ca *eric.charton@bnc.ca

system. We define an optimization function that maximizes both the accuracy and number of intents that can be responded to directly. The goal of this process is to come up with a recommendation of the threshold value on a statistical basis rather than intuition. We conducted our experiments on two question-answering datasets that we obtained from our live dialogue systems. We show that our proposed method improves the accuracy of intent prediction of directly answered queries in both cases by 3-14% while maximizing the number of accurately answered intents compared to the intuitively-set threshold of 70%.

The rest of the paper is arranged as follows: Section 2 outlines related work. Section 3 describes our datasets, followed by the experimental setup in Section 4. We discuss our results in Section 5. Finally, Section 6 concludes our work and discusses future avenues.

2. Related Work

Conversational information seeking has been receiving increasing interest in various studies recently. Many of these studies analyzed real human-human interactions to better understand conversational flow and behavior [6, 7]. The goal is often to understand and reflect those behaviors in human-conversational AI interactions. Clarification questions [8] and turn-taking [9, 10] are some of the recurrent topics in that area. Some dialogue systems are built to engage in a conversation with a user and are known as companion dialogue systems or chatter bots [11]. On the other hand, direct answers (when no further clarifications are required) is a less researched topic despite its prevalence in a natural conversation. In this case, a query or natural language question receives an answer without requiring further information or clarifications from the user [2]. Generally speaking, dialogue systems will provide a user with a direct answer if the confidence of a predicted intent is greater than a certain response threshold. Yu et al. proposed a method for threshold optimization which uses a Markov Decision Process (MDP) combined with reinforcement learning [12] where the threshold is adjusted and re-adjusted as more interactions are performed. In this process, the number of training instances and the range of thresholds constitutes the state space in the optimization problem. This method relies on deploying the dialogue system online and requires a large amount of interactions in order to find the optimal threshold. The question that remains open is how to set the threshold in the earlier stages of a dialogue system.

In the context of a question-answering task in Information Retrieval (IR), Fleischman et al. propose to optimize the classification threshold such as to maximize the recall of the question answering system [13]. Another proposed approach is to optimize the threshold for each individual question thus focusing on a fine-grained gain on a question level [14]. The main difference between IR question answering task and dialogue systems is that the former does not take into account the number of intents (the variety of topics the dialogue system can support). Thus, instead of relying on intuition on what is deemed to be a good threshold, we view this as an optimization problem, where there is a trade-off between accuracy and the number of intents in a dialogue system. We define an optimization function that maximize both accuracy and number of intents that can be responded to directly. Our method has the advantage of being applicable in the scarcity of logs during the initial stages of a dialogue system and can serve as a guideline to initially set the threshold level in an informed fashion. Our method can also be re-used throughout the dialogue system life-cycle to re-adjust this threshold as more logs are available and more intents are covered.

3. Datasets

We conducted our experiments on two datasets that are based on logs of interactions of real users with our corporate conversational AI systems. The main differences between the two datasets are the domain and the volume of intents covered as explained below. The

interactions were in the context of customer support question answering scenario and can be divided into two types: first, interactions on general queries, such as *"how do I change my password?"* and *"open a new account"*. These interactions originate from a dialogue system wherein users had not signed into their profiles or accounts. We refer to this type as ***pre-login***. The second type is what we refer to as ***post-login***[2], which, as the name indicates, is based on interactions with another dialogue system wherein clients had identified their credentials into the transactional platform. In this case, the intents are more specific such as *"visualising the details of a certain transaction"* and *"how to perform an online payment"*.

The pre-login dataset comprises of 900 samples that were randomly pooled from logs of real user interactions that took place throughout October 2020. In all those interactions, the confidence of the predicted intent was above 50%. We define an interaction as a query-response pair, wherein the query was asked by the user, and response was given by the dialogue system based on the predicted intent of the query. Those interactions were then annotated manually to determine whether the predicted intent was correct. The vast majority (around 87%) of those interactions were in French, and only 115 interactions were in English. The annotated interactions cover 23 unique predicted intents out of 32 total intents covered by the dialogue system. The annotated interactions were then split into three disjoint subsets of 300 each. We will refer to these partitions as **PRE-1**, **PRE-2**, and **PRE-3**.

The post-login dataset also consists of 900 randomly sampled logs of real user interactions that occurred during the month of November in 2020. Similar to pre-login, those interactions were manually annotated for intent prediction correctness and, as mentioned above, the confidence level of intent prediction was above 50%. Once again, the post-login interactions were majorly in French with only 175 interactions in English. The number of unique intents predicted in this set is 152, which reflects a richer coverage of topics in those interactions as this dialogue system covers a whopping 271 intents. Those 900 annotated interactions were then split into three disjoint partitions of equal sizes (300 hundred each). We refer to these partitions as **POST-1**, **POST-2**, and **POST-3**.

All the interactions described above are completely disjoint from the training samples of both dialogue systems, as these systems were already trained prior to those interactions.

4. Methodology

Our dialogue systems are based on the Rasa framework. The pipeline consists of the following components: the first one is a pre-processor, which performs several Natural Language Processing (NLP) steps such as tokenization and featurization of the queries to obtain sparse representations at word and character levels. The second component is Rasa's own intent classifier DIET [15] with a Natural Language Understanding (NLU) model that we trained for 200 epochs each. The third component in the pipeline is a disambiguator which is activated whenever the intent classification confidence level is lower than a certain threshold, 70% in our experiments. This disambiguator requires the user to confirm whether the bot *understood* the intent correctly or not. This disambiguation component resolves intents classified with lower confidence further by suggestion pertinent alternative formulations of the user query. In the case of high enough confidence, a direct answer is given without the need to activate the disambiguation component. It is this threshold that we aim to optimize using our proposed method.

The pre-login dialogue system has a fairly narrow domain that concerns mostly login problems, such as forgotten credentials, wrong email or password. It was trained with 708

[2]Note that even though the user is authenticated, their interactions with the dialogue system are anonymized and any potentially personal or sensitive information was masked or removed for the purpose of this work

examples in English, and 1053 examples in French. The post-login dialogue system covers 271 intents, and was trained with 3955 examples in French and 3112 examples in English.

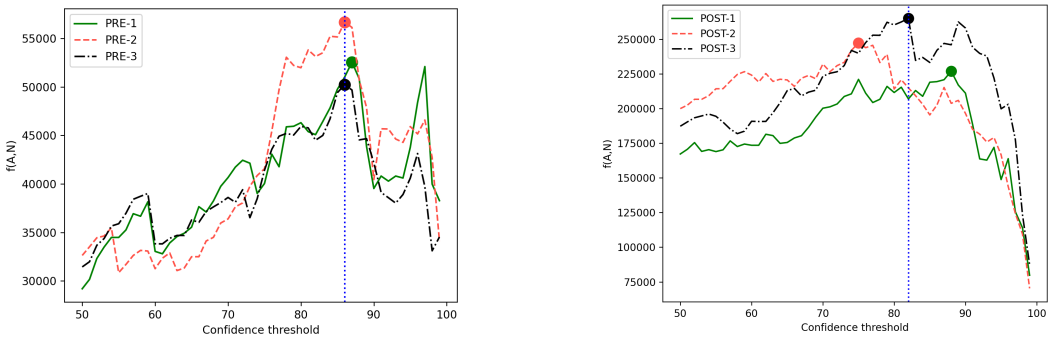
We define the accuracy as the classifier's ability to distinguish between the intents, which translates into the bot's ability to provide a relevant response. For a given sample of a dialogue system, we pose the following optimization problem: both the accuracy and number of intents must be maximized at a certain confidence threshold. As the threshold increases, there is a trade-off between accuracy and the number of intents, where accuracy increases, but the number of intents decreases. Thus, we propose to maximize the following equation:

$$\begin{aligned} &\underset{A,N}{\text{maximize}} && \mathrm{A}^2 \times \mathrm{N} \\ &\text{subject to} && A \in [0, 100] \quad \textit{and} \quad N \in \mathbb{Z}^+ \end{aligned} \tag{4.1}$$

Where A is the accuracy and N is the number of intents. This creates a curve with a global maximum. The threshold at which the global maximum occurs is thus the optimal confidence threshold for a given sample. Each sample is likely to yield a different optimal threshold. Therefore, to determine the confidence threshold for the entire system, the optimization process described above is repeated three times, once on each of the partitions of the datasets. Once all three optimal thresholds are recorded, one for each sample, the arithmetic mean of the three thresholds is considered to be the optimal direct answer threshold for the entire dialogue system. To summarize, for the pre-login system, we perform three iterations, one for each of the partitions PRE-1, PRE-2, and PRE-3. The recommended threshold is the arithmetic mean of the threshold that maximizes each partitions. Similarly, the threshold computed over the runs POST-1, POST-2, and POST-3 (refer to Figures 1a and 1b for optimization curves).

5. **Results and Discussion**

In this section, we present our results and observations for the pre- and post-login datasets compared to their respective baseline systems (threshold 70%). A notable observation is



(a) *(b)*

Figure 1. Optimization curves for pre-login (left) and post-login (right) using our proposed optimization function. The x-axis represents the possible confidence threshold%, and the y-axis represents the optimization function applied on the intent prediction accuracy and the number of intents, as described in 4.1. The maxima are indicated with filled-circles. The vertical dotted line is the arithmetic mean of these maxima, which represent the recommended thresholds

Table 1. Direct answer threshold optimization for pre-login (left) and post-login (right) using our proposed optimization function. The systems marked with * and † represent the baselines for pre-login and post-login systems respectively with a confidence threshold of 70% . *N* represents the number of intents covered in each case

Run	Threshold	Accuracy	*N*
PRE-1	70%*	58.21%	12
	87%	72.50%	10
PRE-2	70%*	52.91%	13
	86%	66.05%	13
PRE-3	70%*	54.50%	13
	86%	64.71%	12

(a)

Run	Threshold	Accuracy	*N*
POST-1	70%†	60.89%	54
	88%	73.50%	42
POST-2	70%†	67.45%	51
	75%	71.05%	49
POST-3	70%†	66.82%	50
	82%	76.73%	45

(b)

that as the confidence threshold increases, the accuracy increases as well, however, the number of intents covered by a direct answer decreases. In Tables 1a and 1b we can see that the optimized threshold using our proposed method increases the overall accuracy for both the pre-login and post-login systems with respect to their baseline counterparts that use an intuitively chosen 70% threshold. In Figures 1a and 1b we can see that the generated curves possess a global maximum for each partition, PRE-1, PRE-2, and PRE-3 for pre-login, and POST-1, POST-2, and POST-3 for post-login. The marked peak in each curve reflects the maximization of both the accuracy and the number of intents, thus projecting the optimal threshold for the partition at hand. As a concrete example, PRE-1 optimal threshold lies at around 86%. The vertical dotted line in each graph indicates the optimal threshold obtained from averaging all three peaks and hence constitutes the recommended direct answer threshold per our method. In the pre-login system, the optimization process increased the accuracy by 10% to 14%, while roughly keeping the same number of intents. As for the post-login system, a notable increase in accuracy is also observed, between 3% and 12%. However, this latter increase is accompanied by a slight decrease in the number of intents covered.

6. **Conclusions and Future work**

In this paper, we propose an optimization methodology to set up the threshold for triggering a direct answer as opposed to responding to the user with clarification questions. Our method estimates an optimal confidence threshold that maximizes both the response accuracy and the number of intents covered. We show that our proposed method improves the accuracy between 3% and 14% (absolute) over the baseline threshold for both datasets. The main advantage of this method is that it can be applied during the early stages of building and releasing the dialogue system as well as during more mature or advanced stages. The optimization can be performed with training data and scarce logs so as to set an optimal direct answer threshold before deploying the system. In the absence or scarcity of real logs, a small test set of queries can be annotated as described in Section 3 to set an initial threshold. Afterwards, the threshold can be assessed and re-optimized iteratively as more interactions become available.

As future steps, we would like to experiment with various optimization functions and study the impact of those variations. For instance, we could explore the impact when accuracy and intent counts are given equal importance, or by accounting for the average number of training samples for all intents into the function. Also, we would like to explore the behavior of the optimization method with datasets of different sizes, such as a much larger number of intents, intents with various numbers of samples, and such combinations. Another possible avenue would be to explore direct answer threshold optimization for each

intent, instead of system-wide optimization. Such an approach might take into consideration several factors, such as the frequency of the intent in real conversations, the number of samples in the initial training set, as well as inter-intent confusions, that is, intents that are often confused by the classifier due to their semantic closeness.

7. Reproducibility

We used Rasa Open Source, version 1.10 which can be installed from https://rasa.com/docs/rasa/installation/. We plan on making the datasets available publicly in the medium term. In the meantime, the datasets can be provided upon request by contacting one of the authors.

References

[1] N. A. Ahmad, M. H. Che, A. Zainal, M. F. Abd Rauf, and Z. Adnan. "Review of chatbots design techniques". In: *International Journal of Computer Applications* 181.8 (2018), pp. 7–10.

[2] M. Potthast, M. Hagen, and B. Stein. "The dilemma of the direct answer". In: *SIGIR Forum*. Vol. 54. 1. 2020.

[3] P. B. Brandtzaeg and A. Følstad. "Why people use chatbots". In: *International conference on internet science*. Springer. 2017, pp. 377–392.

[4] A. Rubin. "Uses and Gratifications. In: Nabi, R.L., Oliver, M.B. (eds.) TheSAGE Handbook of Media Processes and Effects". In: *Decis. Sci.* 35 (2009), pp. 147–159.

[5] E. Luger and A. Sellen. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 5286–5297. ISBN: 9781450333627.

[6] J. Dalton, C. Xiong, V. Kumar, and J. Callan. "CAsT-19: A Dataset for Conversational Information Seeking". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1985–1988. ISBN: 9781450380164.

[7] A. Czyzewski, J. Dalton, and A. Leuski. "Agent Dialogue: A Platform for Conversational Information Seeking Experimentation". In: New York, NY, USA: Association for Computing Machinery, 2020, pp. 2121–2124. ISBN: 9781450380164.

[8] P. Braslavski, D. Savenkov, E. Agichtein, and A. Dubatovka. "What Do You Mean Exactly? Analyzing Clarification Questions in CQA". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. CHIIR '17. Oslo, Norway: Association for Computing Machinery, 2017, pp. 345–348. ISBN: 9781450346771.

[9] B. Galitsky and D. Ilvovsky. "On a Chat Bot Finding Answers with Optimal Rhetoric Representation". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., 2017, pp. 253–259.

[10] H. Zamani and N. Craswell. "Macaw: An Extensible Conversational Information Seeking Platform". In: Association for Computing Machinery, 2020, pp. 2193–2196. ISBN: 9781450380164.

[11] N. Webb, D. Benyon, J. Bradley, P. Hansen, and O. Mival. "Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation." In: *LREC*. 2010.

[12] Y. Yu, A. Eshghi, and O. Lemon. "Training an adaptive dialogue policy for interactive learning of visually grounded word meanings". In: *arXiv preprint arXiv:1709.10426* (2017).

[13] M. Fleischman, E. Hovy, and A. Echihabi. "Offline strategies for online question answering: Answering questions before they are asked". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 2003, pp. 1–7.

[14] R. C. Wang, N. Schlaefer, W. Cohen, and E. Nyberg. "Automatic set expansion for list question answering". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008, pp. 947–954.

[15] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol. "DIET: Lightweight Language Understanding for Dialogue Systems". In: *arXiv preprint arXiv:2004.09936* (2020).