5 Results and Discussion

In this section, we present our results and observations. We begin by analyzing the results obtained from the HOUSE dataset:

The dialogue system directly answers the user's question in more than 40% of the interactions. Roughly half of the interactions entered later stages of the clarification pipeline. The remaining 10% is composed of aborted conversations or the user jumping to another question rather than acting on the confirmation or suggestions.

The Confirmation stage allows the dialogue system to give an answer to the the user (positive confirmation) for 29% of the discussions that go through the clarification process. This 10% of the total number of interactions would not have been responded in a dialogue system in the absence of the multi-stage clarification process and would have ended with a fallback mechanism. Asking for confirmation allows to answer the client's needs without taking the risk of harming the user experience by giving a wrong answer.

In case of a negative answer in the Confirmation stage, the dialogue system will enter the Suggestions stage to propose to the user several possible intents. This mechanism allows to identify the correct intent in 3,5% of the total interactions or 10% of the interactions that go through the clarification process.

figure 2 depicts the distribution of user interactions with the dialogue system across the various stages of our dialogue system based on the HOUSE dataset.

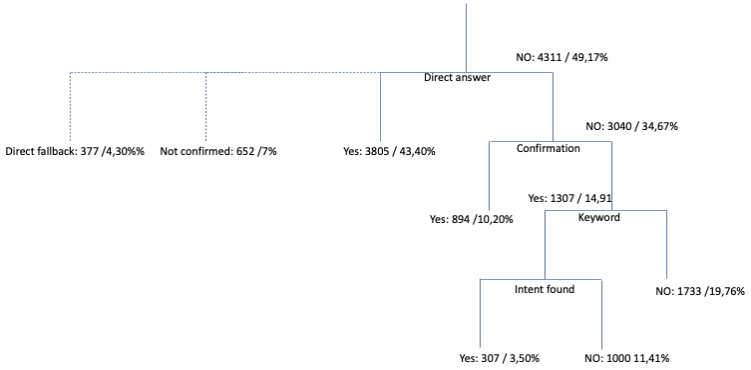

Figure 2: Overall interactions volumes and percentages among the various step of our dialogue system

Our experiments show that our proposed clarification framework allows:

1. To give a correct answer to the user in 10% of the interactions even if the confidence was lower than the threshold. Those 10% would have ended with a fallback mechanism in a classical dialogue system
2. To give the correct answer to the user when the predicted intent is wrong in 3,5% of our experiments by suggesting pertinent canonical formulations of the user query.

Since the data is not annotated, we do not study the impact of direct answer threshold optimization. However a quick analysis shows that two third of the intent predictions where the model confidence was just bellow the threshold (threshold - 0.1) were not confirmed by the user during the clarification stage of our clarification pipeline. We conduct proper comparison with threshold optimisation in the second experiment.

In the second experiment, on SCOPE, we compare the performance of two dialogue systems. One with our proposed clarification pipeline and one with a simple fallback procedure. With the fallback mechanism, the dialogue system does not give answer if its confidence is below a certain threshold. Both dialogue systems have been trained using Rasa DIET classifier with 100 epochs and they share the same NLU engine.

Our clarification pipeline uses keywords to suggest new answers if the user responds negatively to the validation stage. In order to find those keywords we perform a TFIDF analysis on the training examples. Then, we select the top five words with highest TFIDF per intent as keywords for our clarification pipeline.

Finally, our vanilla dialogue system can benefit from the use of threshold optimization. We optimize to find the best possible threshold in order to maximize the number of correctly answered questions. In order to do so we look at the dialogue system performance on the validation set. We select the threshold that maximize the number of good responses to the user queries.

We start by comparing the performance using 0.75 as fallback threshold for the simple vanilla dialogue system. For our improved dialog engine, we also use 0.75 as threshold to trigger the clarification pipeline and 0.3 for direct fallback.

The dialogue system with the clarification component can answer directly 257 (86%) of the user queries with an accuracy for direct answers of 94%. In total 80% of the queries get a correct direct answer. For 3 queries (1%) the dialogue system was confused and couldn't answer the user question (confidence < 0.3). At the CONFIRMATION stage, 45% (18) of the queries have a confirmed intent. For the remaining 22 questions, the dialogue system is able to propose alternative answers through SUGGESTION in 95% (21/22) of the cases. The correct answer is among those propositions in 90% (19/21) of the cases. In total our dialogue system is able to answer the client needs for 93% (277/300) of the interactions, with only 5% (16/300) of wrong answers and 2% (7/300) of fallback.

The dialogue system with simple fallback mechanism and a non optimized threshold set to 0.75 give a correct answer in 80% (241/300) of the interactions, a wrong direct answer in 4% of the cases and the conversation ends with a fallback in 14% of the interactions.

We select the fallback threshold that maximize the number of correct answers given to the user. The selected threshold is 0.35. With such a threshold, the dialogue system gets the following performances: good answers: 86% (259/300), bad answers: 13% (38/300) and fallback: 1% (3/300).

Finally we compare the models performances in term of F1 score: the harmonic mean between precision and recall. The dialogue system with a non-optimized threshold and the dialogue system with our disambiguation component get the same macro-F1 score. By lowering the threshold, the precision of the dialogue system with optimized threshold decrease which leads to lower results in term of macro-F1 score. Regarding the micro-F1, our proposed method get the best results.

	Simple fallback mechanism	Optimized fallback mechanism	clarification mechanism
Good answers	80.3%	86%	92.3%
Bad answers	5.3%	13%	5.3%
Fallback	14.3%	1%	2.3%
macro-F1	0.64	0.51	0.64
micro-F1	0.8	0.86	0.92

Table 2: performances comparison between three dialogue systems: with a simple fallback mechanism, fallback mechanism with optimized threshold and our proposed clarification component

Our clarification pipeline allows us to increase the performance in answering the client's needs by 15% with regard to the vanilla dialogue system and 7% compared to the dialogue system with optimized threshold. The later suffers an increase in the number of incorrect answers whereas our dialogue system can achieve this performance without increasing the number of bad responses.

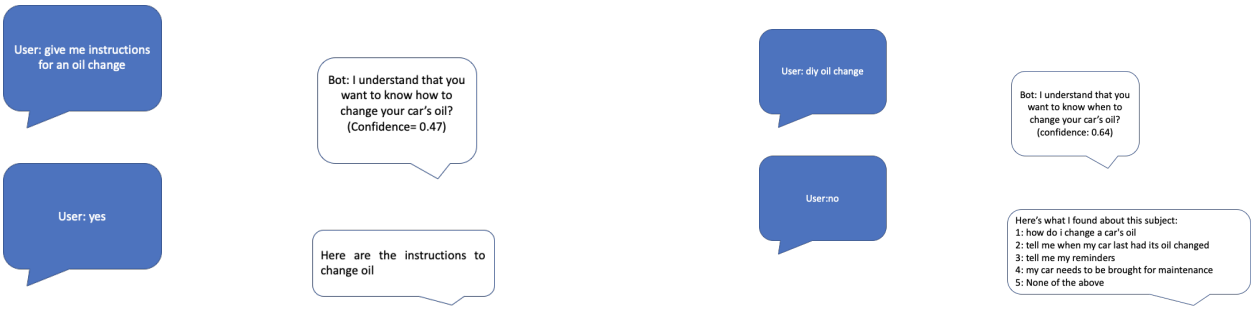

Figure 3: Example of our clarification pipeline for two relatively close intents

Figure 3 illustrates our proposed clarification pipeline in a situation where two intents are very close semantically . We believe our proposed method can prevent the user from the frustration of not being able to find the precise answer to his questions but rather being redirected to something similar.

6 Conclusions and Future work

In paper we propose a multi-stage clarification framework. We show that our proposed framework improves the performance of the dialogue systems. This in turn improves the user experience as relevant answers are given and

clarification is triggered only when needed. This framework reduces the risk of providing hasty, inaccurate answers to the user. When unsure of the user's intent, the dialogue system prompts for confirmation or suggests possible formulations without being unnecessarily highly inquisitive. Our method is simpler than related work on clarification question generation and ranking and is relatively straightforward to deploy and monitor without the need of extra data or model. We conducted our evaluations on two datasets. On the publicly-available in-scope out-of-scope [17] dataset our proposed clarification pipeline allow us to increase the performance in answering the client's needs by 15% with regard to a baseline dialogue system. As a future direction, we will explore click bias and patterns on the interaction with the dialogue system, how the results might differ by device, conversation length/stage, and order of the suggestions. Further work may also include further customer-specific answers and clarification questions based on click behaviour and implicit feedback or using external info held on the client (bank account, previous transactions) to propose better answers and clarification.

References

[1] Nahdatul Akma Ahmad, Mohamad Hafiz Che, Azaliza Zainal, Muhammad Fairuz Abd Rauf, and Zuraidy Adnan. Review of chatbots design techniques. *International Journal of Computer Applications*, 181(8):7–10, 2018.

[2] Asbjørn Følstad and Petter Bae Brandtzaeg. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience*, 5(1):1–14, 2020.

[3] Kathleen R McKeown. Paraphrasing using given and new information in a question-answer system. *Technical Reports (CIS)*, page 723, 1980.

[4] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428, 2020.

[5] W. Bruce Croft. The importance of interaction for information retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1–2, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Nicholas J Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, 9(3):379–395, 1995.

[7] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina 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, page 345–348, New York, NY, USA, 2017. Association for Computing Machinery.

[8] Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, and Haiqing Chen. Iart: Intent-aware response ranking with transformers in information-seeking conversation systems. In *Proceedings of The Web Conference 2020*, WWW '20, page 2592–2598, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Vaibhav Kumar, Vikas Raunak, and Jamie Callan. Ranking clarification questions via natural language inference. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2093–2096, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N Bennett, Nick Craswell, and Susan T Dumais. Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1181–1190, 2020.

[13] Sudha Rao and Hal Daumé III. Answer-based adversarial training for generating clarification questions. *arXiv preprint arXiv:1904.02281*, 2019.

[14] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. User engagement prediction for clarification in search. *arXiv preprint arXiv:2102.04163*, 2021.