CL-QR: Cross-Lingual Enhanced Query Reformulation for Multi-lingual Conversational AI Agents

Zhongkai Sun[1] **Zhengyang Zhao**[1] **Sixing Lu**[1] **Chengyuan Ma**[1]
Xiaohu Liu[1] **Xing Fan**[1] **Wei Shen**[1] **Chenlei Guo**[1]
[1]Amazon Alexa AI
{zhongkas,zzhengya,cynthilu,mchengyu,derecliu,fanxing,sawyersw,guochenl}@amazon.com

Abstract

The growing popularity of conversational AI agents such as Alexa, Google Assistant, and Siri rely on accurate spoken language comprehension. The query reformulation (QR) method, which reformulates defective user queries, has been broadly adopted to mitigate the challenges posed by understanding user's intent from imperfect spoken recognition result. However, due to the scarcity of non-English QR labels, providing high-quality QR for non-English users still remains a challenge. This work proposes a novel cross-lingual QR framework, CL-QR, to leverage the abundant reformulation resources in English to improve non-English QR performance. The proposed work also proposes a **M**odule-wise **M**utually-supervised **F**eedback learning (MMF) algorithm to enable the continually self-improving of the CL-QR, which alleviates the lack of cross-lingual QR training data and enhances the delivery of high-quality reformulations learned in English for multilingual queries. Both offline evaluation and online A/B testing demonstrates the effectiveness of the proposed method.

1 Introduction

Conversational AI agents like Alexa, Siri, and Google Assistant are becoming ubiquitous today. However, the inaccurate interpretation of users' queries is a critical issue, in which considerable number of interpretation failures come from the frictions induced by either Automatic Speech Recognition (ASR) error, semantic ambiguity, or defective user expressions. Query reformulation (QR) techniques (Sun et al., 2022; Hao et al., 2022; Cho et al., 2021) have been widely adopted to improve the comprehension of user intentions for AI agents. For instance, for the query "*play old town load*" which contains an ASR error, QR system is employed to fix it to the correct one "*play old town road*"; and for the query "*where is danube*" which is incomplete, QR system can reformulate it to "*where is the danube river*". Building production-level QR systems requires substantial amounts of user-agent interactive data (e.g., user queries, implicit user feedback, and user rephrases) (Hao et al., 2022; Ponnusamy et al., 2020, 2022). The QR systems, however, struggle to provide high-quality reformulations for non-English queries due to the limited number of non-English users/traffic.

To address this challenge, this work proposes a novel **C**ross-**L**ingual **Q**uery **R**eformulation (CL-QR) system that effectively leverages the abundant query reformulation resources in English to generate superior reformulations for non-English queries. Figure 1 demonstrates several cross-lingual reformulation examples that can be achieved by CL-QR.

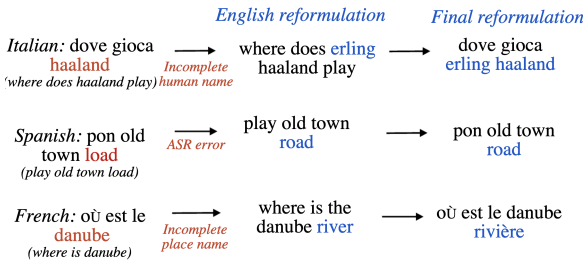


Figure 1: Examples demonstrating how CL-QR works. The defective non-English queries are first mapped to English to get the English reformulations, and then transferred to final reformulations in the original language.

CL-QR comprises three major components: *cross-lingual reformulation extraction*, *cross-lingual reformulation plausibility detection*, and *back translation*. The *cross-lingual reformulation extraction* aims at extracting all potential English reformulations for a defective non-English query; the *cross-lingual reformulation plausibility detection* predicts the plausibility score for each cross-lingual QR pair and the score is used to select the most suitable English reformulation; the *back translation* is used to translate the most suitable English reformulation back to the original language.

To address the challenge of insufficient CL train-

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 423–431
December 6-10, 2023

ing data, this paper proposes a novel **M**odule-wise **M**utually-supervised **F**eedback learning (MMF) algorithm. Specifically, the *plausibility detection module* and the *back-translation module* can be trained in a mutually reinforcing manner, in which each module's output can be used as a weakly feedback label to supervise the other module's training. In this way, the CL-QR can be initially trained on a small set of golden cross-lingual QR pairs, and then continually improves itself using readily available large-scale mono-lingual non-English QR pairs. The offline evaluation and online A/B test on Spanish, Italian, and French demonstrate the efficacy of the proposed CL-QR.

2 Related Work

Query reformulation: Query reformulation aims to correct the frictions in user utterances in voice controlled AI agent, and has been widely studied. For example, search/retrieval based methods are used to find the ideal reformulation from a collection of successful utterance (Fan et al., 2021; Cho et al., 2021; Sun et al., 2022; Naresh et al., 2022), generation methods are applied to generate the reformulation directly (Hao et al., 2022; Yu et al., 2020), and Markov chain models coverage query reformulation patterns based on the collaborative filtering mechanism (Ponnusamy et al., 2020, 2022). Although these methods have achieved great performance on English conversational AI system, their performances on non-English languages are suboptimal because of the data sparsity challenges.

Cross-Lingual Knowledge Transferring: Pretrained multilingual language models like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mBART (Liu et al., 2020) have opened the doors for cross-lingual transfer learning, particularly focusing on transferring knowledge from resource-rich languages to resource-scarce languages. Such idea has been widely studied and applied in various tasks such as question answering (Roy et al., 2020; Asai et al., 2021), keyphrase generation (Gao et al., 2022), e-commerce product search (Ahuja et al., 2020; Zhang and Misra, 2022), named entity recognition (NER) (Liu et al., 2021; Zhou et al., 2022), natural language understanding (NLU) (Xu et al., 2020; Abujabal et al., 2021), etc. In this work, we adopt the idea for QR task, based on the fact that many QR patterns are shareable across languages (e.g. fixing the ASR error in entities, completing a user query, etc.).

3 Methodology

This section describes CL-QR in details. Given a defective query in language-X (non-English) as input (X_Q), our ultimate goal is to find a proper non-defective reformulation of it (X_R). To achieve this, the CL-QR framework takes 3 steps as illustrated in Figure 2a: 1) Map X_Q to the corresponding English query EN_Q (usually also defective) through cross-lingual retrieval or translation, and then perform QR in English to get EN_Q's non-defective reformulations EN_R; 2) Use the *plausibility detection* module to select the most appropriate EN_R for X_Q, from the multiple candidates from step 1); 3) Finally, the language-X reformulation X_R is achieved by translating the selected EN_R back into language-X.

3.1 Cross-lingual Reformulation Extraction

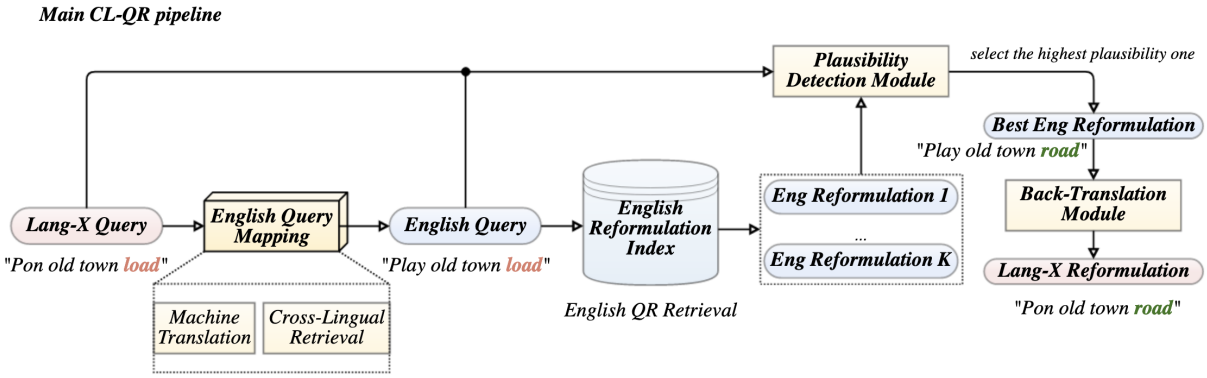
As shown in Figure 2a, the input query X_Q is first mapped to a query in English (EN_Q) with the same semantics so that the English query-reformulation method can be applied. This cross-lingual mapping can be obtained through different approaches, such as machine-translation and cross-lingual semantic retrieval. For one X_Q, we can keep N candidates for EN_Q to increase the recall ($|set(EN_Q)| = N$).

Next, English reformulations (EN_R) for each EN_Q can be obtained through the English QR system. Here, we use a dual-encoder model to retrieve EN_R from the English reformulation index (a collection of high quality English queries). Top-K reformulations are retrieved for each EN_Q, therefore for one original input X_Q, the corresponding English reformulation set $set(EN_R)$ can be established with the size of $N \times K$.

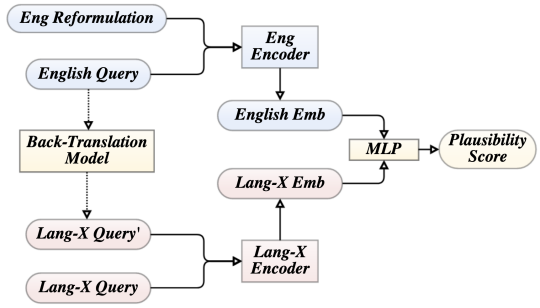
3.2 Final Reformulation from Plausibility Detection and Back Translation

To select the most appropriate English reformulation for X_Q from the $set(EN_R)$, a plausibility detection module is defined to measure the plausibility of each cross-lingual QR pair. The plausibility refers to the degree of semantic congruity between the English reformulation and the language-X query (Relevant examples can be found in Appendix A.2.1). A language-X semantic encoder and an English semantic encoder are leveraged in this module. Figure 2b illustrates the details of the plausibility detection module.

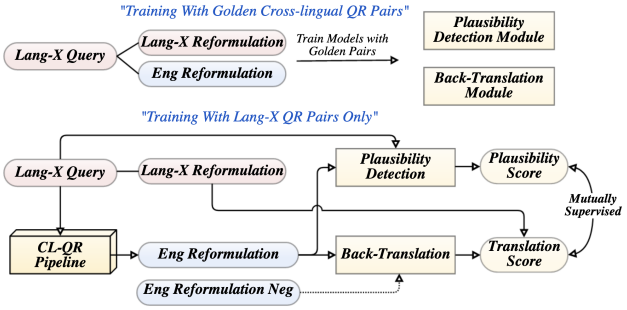
For a specific set of *raw query, EN query, and EN reformulation* (<X_Q, EN_Q, EN_R>), the EN



(a) The overview of the proposed CL-QR framework



(b) Plausibility-Detection Module



(c) Training pipelines. A mutually supervised strategy is designed to enhance self training based on module wise feedback.

Figure 2: Illustration of the CL-QR framework (a), plausibility detection module (b) and training pipeline (c). In figure (b) and (c), dashed lines indicate steps without gradient computation/update.

query EN_Q is first back-translated to language X to obtain the back-translated query $X_Q^{'}$. Next, $X_Q^{'}$ and the raw query X_Q are input to a language-X encoder to indirectly encode the semantic consistency between the English query EN_Q and the raw query X_Q. Besides, EN_Q and its reformulation EN_R are also input to the English encoder to encode the correctness of the *EN* reformulation. Finally, the plausibility score is predicted using a Multi-Layer Perceptron (MLP) that takes the output of the two encoders as the input.

Once the plausibility scores are calculated for each set of $<X_Q, EN_Q, EN_R>$, the best English reformulation EN_R^{best} can be obtained by selecting the EN_R that with the highest plausibility score. After that, a back-translation model is utilized to translate the EN_R^{best} to the final reformulation in language X.

3.3 Training Plausibility Detection and Back-Translation Modules

Training the plausibility detection module and the back-translation module relies on golden cross-lingual QR pairs (parallel data of $<X_Q, X_R, EN_R>$), which is hard to be obtained. However, on the other hand, the large-scale mono-lingual QR pairs ($<X_Q, X_R>$) in language X can be easily obtained but cannot be directly leveraged due to the lack of English counterparts.

To address the aforementioned challenges, instead of training the plausibility detection and the back-translation modules individually, a novel **M**odule-wise **M**utually-supervised **F**eedback learning (MMF) training algorithm is designed to train the two modules on the mono-lingual QR data in a mutually supervised manner. As shown in Figure 2c, at each training step: 1) the plausibility score $score_p$ is firstly calculated using the method described in section 3.2; 2) the English reformulation candidate EN_R together with a pre-selected negative reformulation EN_{Rneg} (which contains different semantic meaning) are first back-translated to $X_R^{'}, X_{Rneg}^{'}$. Then the language-X encoder E_X is applied to calculate the cos-similarities between the back-translations and the language-X reformulation X_R, i.e., $cos(X_R^{'}, X_R)$ and $cos(X_{Rneg}^{'}, X_R)$. A reward r is then calculated as:

$$r = max(0, cos(X_R^{'}, X_R) - cos(X_{Rneg}^{'}, X_R))$$

Such reward can measure the effectiveness of the current back-translation model, wherein a higher value signifies superior ability of the back-translation model to differentiate between the reformulation and its negative counterparts. Then the back-translation loss L_{trans} can be weighted using the reward r:

$$L_{trans} = BackTransModel(EN_R, X_R)$$
$$L^*_{trans} = r * L_{trans}$$

The plausibility $score_p$, weighted translation loss L^*_{trans} can be used in a self-supervised training based on the module-wise feedback iteratively. Specifically, if the $score_p$ is higher than a threshold T, then the back-translation model will be updated using L^*_{trans}; meanwhile, if the $score_p$ is not aligned with the reward r, the plausibility model will be updated accordingly. Algorithm 1 illustrates the details. During the training, the plausibility module and back-translation module can be initially trained on a small set of golden cross-lingual QR data, and then fine-tuned on the large set of language-X QR data (without English reformulation label) using the training strategy above.

Algorithm 1 MMF Training Algorithm

Input: Batches of (X_Q, X_R, EN_Q, EN_R, $X^{'}_Q$, $X^{'}_R$, $X^{'}_{Rneg}$)
Set up Plausibility Module M_P, Back Translation Module M_B, Reward Module M_R
Set plausibility upper threshold T $\in (0, 1)$, lower threshold N $\in (0, 1)$
for i in Num_{batch} **do**
  $score_p \leftarrow M_P(X^i_Q, X^{'i}_Q, EN^i_Q, EN^i_R)$
  $L_{trans} \leftarrow M_B(EN^i_R, X^i_R)$
  $r \leftarrow M_R(X^{'i}_R, X^{'i}_{Rneg}, X^i_R)$
  $L^*_{trans} \leftarrow r * L_{trans}$
  Freeze M_P
  if $score_p > T$ **then**
    Update M_B using L^*_{trans}
  end if
  Freeze M_B
  if $score_p > T$ AND $r == 0$ **then**
    L_p = CrossEntropy($score_p$, 0)
  else if $score_P < N$ AND $X^{'}_R$ == X_R **then**
    L_p = CrossEntropy($score_p$, 1)
  end if
  Update M_P using L_p
end for

4 Experiments

4.1 Dataset

The CL-QR framework is built and evaluated in three non-English languages: Spanish, French, and Italian. The training and offline test data are collected from historical user traffic[1]. For training, two training sets are built on each language:

Mono-lingual QR training set: each sample consists of mono-lingual QR pairs <X_Q, X_R> in language-X, which have been verified to be successful reformulation pairs in the historical traffic. This set is relatively large.

Cross-lingual QR training set: each sample consists of a golden cross-lingual QR pair <X_Q, X_R, EN_R>. This set is small and is used for the initial training as described in Section 3.3.

Additionally, for each language, a **reformulation index** and a **test set** is built. Here, "reformulation index" refer to a collection of non-defective queries, from which the search-based QR systems can retrieve the ideal reformulation given an defective input. The test set for offline evaluation consist of QR pairs <X_Q, X_R>, similar with the mono-lingual QR training set. More details of the data building can be found in the appendix A.1, Table 1 presents the dataset statistics.

Lang	Test	Cross-lingual Train	Mono-lingual Train	Index
Spanish	11k	30k	3M	7M
French	38k	40k	2M	4M
Italian	12k	36k	2.5M	5M
English	N/A	N/A	N/A	12M

Table 1: Train, Test, and Index Statistics

4.2 Experiment Setup

The overall performance of CL-QR is evaluated by two metrics: 1) the reformulation precision, i.e., if the generated reformulation exactly matches the ground truth; and 2) the average BLEU score between the ground truth and generated reformulation. Besides, the plausibility detection and translation loss on the test set are used to evaluate the performance of the self-supervised interactive training of the plausibility and back-translation modules. For the online A/B test, the conversation-level defect rate is calculated using (Gupta et al., 2021) for control / treatment sets, which is used to evaluate the effectiveness in real-world user experience.

[1]Note that all data used in this paper has been de-identified so that no user information is remained

Model Setup As introduced previously, the CL-QR framework comprises 3 components: *CL reformulation extraction*, *Plausibility detection*, and *Back translation*. The following is the setup for each module.

CL reformulation extraction: the language-X input query X_Q is first mapped to a English query EN_Q in two ways: using the mBART (Tang et al., 2020) to do translation, or to do cross-lingual retrieval with a CL encoder fine-tuned on LABSE (Feng et al., 2022). Then, to extract EN_Q's reformulation EN_R, a dual-encoder model (Karpukhin et al., 2020) trained on top of XLM-R (Conneau et al., 2020) is used. The retrieval count K is set as 3. All components in the *CL reformulation extraction* module are trained beforehand to obtain optimized functionality, and will not be updated during the training process of the *plausibility detection* and *back translation* modules.

Plausibility detection and *Back translation*: as discussed in Section 3.3, we leveraged mBART-large as the *back translation* model, and XLM-R encoders followed by a three-layer MLP (whose hidden-size is 128) as the *plausibility detection* model. The learning rate is set as $5e-5$ and the AdamW is used as the optimizer. The plausibility upper threshold *T* in Algorithm 1 is set as 0.7, and the lower threshold *N* is set as 0.3. All the trainings are conducted on eight NVIDIA Tesla V100 GPUs, with epoch number set as 5.

Baselines The following SOTA QR methods as well as methods leveraging recent large language models (LLM) are used as our baslines:

Mono-retrieval: a dual-encoder model (Karpukhin et al., 2020) is trained on top of XLM-R and used as the monolingual-retrieval baseline, which conducts the X_Q-X_R retrieval directly from the index of Language-X.

MUFS-QR: the search-based QR method UFS-QR (Fan et al., 2021) composed of a retrieval layer and a ranking layer, is extended to multi-lingual version MUFS-QR.

MCGF: the generation-based QR model CGF (Hao et al., 2022) is extended to multi-lingual version MCGF.

Vicuna-13B-zs/fs: zero-shot (zs) and few-shot (fs) experiments conducted on the recent released Vicuna-13B LLM (Chiang et al., 2023). [2]

[2]Note that due to data restriction policies, the most powerful Chat-GPT and GPT4 cannot be applied to our data for performance comparison.

4.3 Offline Experiment Results

4.3.1 Overall Performance

The overall QR performance on the Spanish, French, and Italian golden QR test set are reported in Table 2 (measured by precision) and 3 (measured by BLEU score). For each language X, the test pairs <X_Q, X_R> can be classified into two groups: "*Lang Index Covered*" indicates the samples whose ground-truth reformulations X_R are contained in the language-X index. Conversely, the "*Lang Index Not-Covered*" refers to data whose X_R are not included in the language-X index. The latter scenario happens frequently in non-English languages because of data scarcity.

Overall, baselines relying on retrieval/search methods (Mono-retrieval and MUFS-QR) demonstrate limited performance on *Lang Index Not-Covered* set, owing to the fact that the golden reformulation is not covered in the index of lang-X. The generation-based method (MCGF) achieves the best performance for *Lang Index Covered* set, but its performance drops sharply on *Lang Index Not-Covered* set. It is because that MCGF is only trained on QR pairs and lacks the ability for cross-lingual knowledge transfer. Besides, for LLM (e.g. Vicuna-13B), zero shot and few shot also have very limited performance for the QR task, demonstrating that such LLMs' limited capability in fixing non-English spoken recognition errors.

In contrast, the proposed CL-QR method is able to achieve the best performance on the *Lang Index Not-Covered* set, and is also effective on the *Lang Index Covered* set (note that the CL-QR model isn't fine-tuned on additional mono-lingual QR tasks, which may limit its performance on the *Lang Index Covered* set). Therefore, the CL-QR is able to achieve the best performance on the overall test set. Appendix A.2.2 shows relevant examples.

4.3.2 Back-translation and Plausibility Detection Performance

The performance of the plausibility detection, the back-translation, and the ablation study for the self-supervised training strategy are shown in Table 4.

We can conclude that the model can achieve a good plausibility detection accuracy when only trained on the small set of golden cross-lingual QR data. However, the back-translation loss remains high due to the limited size of the data. Besides, when trained on the mono-lingual QR data only, the model's back-translation loss is decreased because

Model	Lang Index Covered			Lang Index Non-Covered			Overall Precision		
	Spanish	French	Italian	Spanish	French	Italian	Spanish	French	Italian
Mono-retrieval	0.0	0.0	0.0	N/A	N/A	N/A	0.0	0.0	0.0
MUFS-QR	+14.0%	+15.2%	+15.6%	N/A	N/A	N/A	+16.7%	+15.4%	+15.0%
MCGF	**+23.2%**	**+26.1%**	**+24.4%**	0.0	0.0	0.0	+50%	+42.3%	+55.0%
CL-QR	+18.6%	+19.6%	+22.2%	**+90.9%**	**+137.5%**	**+60.0%**	**+88.9%**	**+53.8%**	**+65.0%**
Vicuna-13B-zs	N/A	N/A	N/A	N/A	N/A	N/A	-69.33%	-74.71%	-80.06%
Vicuna-13B-fs	N/A	N/A	N/A	N/A	N/A	N/A	-32.11%	-37.30%	-50.10%

Table 2: Overall query reformulation performance measured by precision. The value 0's represent the base precision and the relative improvement of each model is reported. "Lang Index Covered" represents the test data whose language-X reformulations exist in the language-X index, while the "Lang Index Not-Covered" is the opposite.

Model	Lang Index Covered			Lang Index Non-Covered			Overall BLEU		
	Spanish	French	Italian	Spanish	French	Italian	Spanish	French	Italian
Mono-retrieval	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MUFS-QR	+7.7%	+4.8%	+0.0%	+7.1%	+3.8%	+23.0%	+5.3%	+4.3%	+10.8%
MCGF	**+13.5%**	**+4.8%**	**+17.0%**	35.7%	34.6%	+28.0%	+23.7%	+10.6%	+21.6%
CL-QR	+9.6%	-3.3%	+1.9%	**+85.7%**	**+96.2%**	**+88.0%**	**+39.5%**	**+36.2%**	**+35.1%**
Vicuna-13B-zs	N/A	N/A	N/A	N/A	N/A	N/A	-14.92%	-21.74%	-23.66%
Vicuna-13B-fs	N/A	N/A	N/A	N/A	N/A	N/A	-2.51%	-6.79%	-11.13%

Table 3: Overall query reformulation performance measured by BLEU score. The BLEU score is calculated between the obtained reformulation the ground truth. The value 0's represent the base BLEU score.

Model	Spanish		French		Italian	
	T_{loss}	P_{acc}	T_{loss}	P_{acc}	T_{loss}	P_{acc}
Without Training	10.44	N/A	9.57	N/A	9.65	N/A
Cross-lingual-Train Only	1.36	77%	1.47	79%	1.51	75%
Mono-lingual-Train Only	0.78	63%	0.89	59%	0.86	60%
Joint Train w/o MMF	0.45	77%	0.42	79%	0.46	75%
Joint Train w/ MMF	**0.33**	**84%**	**0.28**	**85%**	**0.34**	**81%**

Table 4: Results of the back-translation loss and the plausibility detection accuracy.

of the large size of the data. However, due to the lack of golden cross-lingual reformulation labels, the model may bias toward generating sub-optimal translations, which also impacts the ability to detect plausibility.

Jointly training on both cross-lingual and mono-lingual QR data (the model is trained on the cross-lingual data first and then trained on the mono-lingual data) without the proposed MMF in Sec. 3.3 can further reduce the back-translation loss. However, the plausibility detection module cannot be updated as there are no labels for the plausibility training. After training with the MMF on both datasets, the model achieves the best performance in reducing translation loss and increasing plausibility accuracy. This result successfully verifies the effectiveness of the proposed training strategy.

4.4 Online A/B Test Results

The online A/B testings were conducted separately for Spanish, French, and Italian product traffic. The performance for the control/treatment groups were measured by calculating the average defect rate at each user-agent interaction session level using (Gupta et al., 2021). The defect rate quantifies the proportion of interactions in which the user's intent is not accurately identified, lower rate corresponds to a superior user experience.

The A/B test results show that, in comparison to the control group, the average defect rate declined by **8.0%**, **7.7%**, and **11.0%** for Spanish, French, and Italian users, respectively. These findings further validate the effectiveness of CL-QR in enhancing the non-English user experience.

5 Conclusion

In this work, a novel Cross-Lingual Query Reformulation (CL-QR) method is proposed to address the limitations of existing QR systems for non-English queries by leveraging large-scale of QR resources in English to generate reformulations for low-resource non-English queries. The CL-QR also includes a novel plausibility detection module to select the best cross-lingual reformulations. Additionally, a module-wise mutually-supervised feedback training strategy is proposed for effective self-training. The proposed CL-QR method has been rigorously evaluated through offline testing and online A/B experiments conducted for Spanish, Italian, and French traffic. The promising results verify the effectiveness of this framework.

References

Abdalghani Abujabal, Claudio Delli Bovi, Sungho Ryu, Turan Gojayev, Fabian Triefenbach, and Yannick Versley. 2021. Continuous model improvement for language understanding with machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, pages 56–62, Online. Association for Computational Linguistics.

Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K. Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In WSDM 2020.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 547–564, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized search-based query rewrite system for conversational ai. In Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 179–188.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xing Fan, Eunah Cho, Xiaojiang Huang, and Edward Guo. 2021. Search based self-learning query rewrite system in conversational ai.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael Lyu. 2022. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1233–1246, Seattle, United States. Association for Computational Linguistics.

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Edward Guo. 2021. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems.

Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Edward Guo, and Gokhan Tur. 2022. Cgf: Constrained generation framework for query rewriting in conversational ai.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5834–5846, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

Niranjan Uma Naresh, Ziyan Jiang, Ankit, Sungjin Lee, Jie Hao, Xing Fan, and Edward Guo. 2022. Pentatron: Personalized context-aware transformer for retrieval-based conversational understanding. In EMNLP 2022.

Pragaash Ponnusamy, Clint Solomon Mathialagan, Gustavo Aguilar, Chengyuan Ma, and Chenlei Guo. 2022. Self-aware feedback-based self-learning in large-scale conversational ai. arXiv preprint arXiv:2205.00029.

Pragaash Ponnusamy, Alireza Roshan Ghias, Chenlei Guo, and Ruhi Sarikaya. 2020. Feedback-based self-learning in large-scale conversational ai agents. Proceedings of the AAAI Conference on Artificial Intelligence, 34(08):13180–13187.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5919–5930, Online. Association for Computational Linguistics.

Zhongkai Sun, Sixing Lu, Chengyuan Ma, Xiaohu Liu, and Chenlei Guo. 2022. Query expansion and entity weighting for query reformulation retrieval in voice assistant systems. arXiv preprint arXiv:2202.13869.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5052–5063, Online. Association for Computational Linguistics.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pages 1933–1936.

Bryan Zhang and Amita Misra. 2022. Machine translation impact in e-commerce multilingual search. In EMNLP 2022.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency training for cross-lingual named entity recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Data Collection

A.1.1 Reformulation Index Collection

In this paper, we use the term "reformulation index" to refer to a collection of non-defective queries, from which the search-based QR system can retrieve the ideal reformulation given an defective input. The reformulation index for each language is established by collecting each language's de-identified successful user queries from the historical traffic. Specifically, filters are applied on queries' frequency, queries' defect rate, etc., to guarantee all queries in the reformulation index are non-defective and can be successfully handled by the voice assistant system. The defect rate of each query is measured by using the model proposed in (Gupta et al., 2021).

A.1.2 QR Data Collection

Mono-lingual query reformulation training data (consists of $<X_Q, X_R>$ pairs) is extracted from de-identified traffic. Specifically, existing query reformulation pairs are extracted from each language's traffic log. Such existed pairs are either from human-annotation or existing mono-lingual QR systems. Filters are applied on reformulation frequency, reduction of defect rate, etc., to guarantee the extracted QR pairs have successfully improved customer experience in the history.

In order to select the cross-lingual QR training data (parallel data of $<X_Q, X_R, EN_R>$), a cross-lingual semantic-matched reformulation table is first established based on either human-annotation or strict machine translation criteria. Such reformulation table is able to link the high-confident reformulations that in different languages. Then, the cross-lingual QR data can be established by linking the reformulations in the mono-lingual QR data with the English reformulations in the semantic-matched reformulation table.

A.2 Case Study

A.2.1 Plausibility Score

In this work, the plausibility score is the measure of semantic alignment that exists between the English reformulation and the language-X query. Table 5 demonstrates the examples of different cross-lingual QR pairs' plausibility scores.

A.2.2 Comparison of Mono-lingual QR and cross-lingual QR

Table 6 demonstrates the comparison of reformulations obtained from mono-lingual reformulation method and CL-QR. The mono-lingual QR, owing to the lack of sufficient QR data, is not able to provide an accurate reformulation.

B Ethical Discussion

The development and implementation of cross-lingual QR for conversational AI agents can improve accessibility and convenience for users. However, the use of these technologies may raise ethical considerations, particularly with regards to privacy

Language	Query	EN Query	EN Reformulation	Score
Italian	dove gioca haaland	where does haaland play	**where does erling haaland play**	**0.86**
Italian	dove gioca haaland	where does haaland play	who is haaland	0.62
Italian	dove gioca haaland	birthday of haaland	when is haaland's birthday	0.51
Spanish	pon old town load	play old town load	**play old town road**	**0.90**
Spanish	pon old town load	listen old town load	old town road	0.82
Spanish	pon old town load	old city story	old city of tacoma	0.47
French	où est le danube	where is the danube	**where is the danube river**	**0.85**
French	où est le danube	where is the danube	how long is danube river	0.68
French	où est le danube	weather in danube	what is the weather in danube	0.34

Table 5: Plausibility score examples for different cross-lingual pairs. The correct reformulation and its score is in bold.

Language	Query	Reformulation from Mono-lingual QR	Reformulation from CL-QR
Italian	dove gioca haaland (where haaland plays)	quanti anni ha harland (how old is harland)	dove gioca erling haaland (where does erling haaland play)
Spanish	pon old town load (play old town load)	pon old town (play old town)	pon old town road (play old town road)
French	où est le danube (where is the danube)	hôtel proche du danube (hotels near the danube)	où est le danube rivière (where is the danube river)

Table 6: Comparisons between outputs from mono-lingual QR method and CL-QR

and data protection. For example, the extraction of cross-lingual reformulation may involve the collection and processing of large amounts of user data. It is important to ensure that the collection and processing of this data is done in a way that protects user privacy and that user data is not misused or mishandled. Additionally, the use of CL-QR may raise concerns about linguistic and cultural bias in the data used to train the model. It is also important to ensure that the method is designed in a way that takes into account the diverse linguistic and cultural backgrounds of the users it serves In this work, all the data used for training / testing are from data with identification information removed. This means that users' personal information have been removed and will not be fed into the models. Besides, the production pipeline also includes several guardrails to filter out inappropriate reformulations, to ensure the cross-lingual reformulation doesn't include any cultural bias.