Query Rewriting under Noise

Romain Delhommais[*]

Abstract

We introduce a simple yet effective pipeline for noise-robust query rewriting under realistic user error. First, we simulate typing mistakes by applying random QWERTY-neighbor swaps, token deletions, and character replacements to clean queries from MS MARCO and Natural Questions (NQ). We then fine-tune a Flan-T5 teacher model on these paired (noisy to clean) examples to learn the denoising rewriting task. Next, we perform sequence-level knowledge distillation: a compact T5-Flan-small student is trained to mimic the teacher's rewritten outputs on the same noisy inputs, using a combined cross-entropy and distillation loss. Qualitative evaluation on noisy queries shows that the distilled student recovers near-teacher rewriting quality. Our results highlight that distillation makes noise-robust query rewriting practical for deployment in resource-constrained settings.

1 Introduction

In modern information retrieval systems, user queries often contain typographical errors, omissions, or other forms of noise that degrade downstream model performance. Such errors may arise from hurried typing on small devices, unfamiliarity with keyboard layouts, or simple human mistakes. As a result, many state-of-the-art retrieval and question answering pipelines, which assume clean inputs, suffer a significant drop in accuracy when confronted with noisy queries.

Previous approaches to noise-robust query rewriting have relied on large sequence-to-sequence models trained directly on noisy-clean pairs or on synthetic noise injection schemes. While these methods can improve robustness, they typically incur high computational and memory costs, making them impractical for deployment in latency-sensitive or resource-constrained settings (e.g. mobile devices or real-time services). Moreover, existing noise models do not always reflect realistic user behavior, limiting the transferability of academic results to production environments.

To address these challenges, we propose a simple yet effective three-stage pipeline for noise-robust query rewriting. First, we design a realistic noise simulator that applies QWERTY-neighbor swaps, token deletions, and character replacements to generate diverse noisy variants of clean queries from MS MARCO and Natural Questions. Second, we fine-tune a Flan-T5 teacher model on these synthetic noisy-clean pairs to learn high-quality denoising rewriting. Finally, we perform sequence-level knowledge distillation to train a compact Flan-T5-small student model to mimic the teacher's outputs, using a combined cross-entropy and distillation loss.

Our experiments demonstrate that the distilled student recovers most of the teacher's rewriting accuracy on held-out noisy queries, while reducing model size and inference latency. These results show that our approach makes noise-robust query rewriting both effective and practical for deployment in environments with strict resource constraints.

[*] Project conducted with Vincent Gimenes; this report is written individually.

2 Related Work

2.1 Query Error Correction

Traditional query error correction methods treat misspellings as a noisy-channel decoding problem, learning a generic edit error model via EM from misspelled/corrected pairs. Brill and Moore [1] propose an improved string-to-string error model estimated in an EM framework, eschewing hand-crafted rules. Kukich [3] surveys non-word detection, isolated-word correction, and context-dependent techniques-including pattern matching, n-grams, and early statistical models. He et al. [7] introduce KSTEM, a sequence-to-edit tagging model mapping noisy Chinese queries into KEEP/REPLACE/SWAP/DELETE/INSERT tags with 2D position encodings and a multi-stage pretraining paradigm for efficient deployment in search engines.

2.2 Query Paraphrasing

Prakash et al. [5] propose a stacked residual LSTM network for paraphrase generation, outperforming vanilla seq2seq on PPDB, WikiAnswers, and MSCOCO benchmarks. Raffel et al. [6] introduce T5, a unified text-to-text transformer that casts all NLP tasks as sequence-to-sequence, enabling effective transfer learning across tasks; while the original work does not specifically target query rewriting, T5 has been widely applied to paraphrasing.

2.3 Contextual Query Reformulation

Yu et al. [8] develop an utterance rewriter using multi-turn dialogue context to resolve coreference and ellipsis, generating contextually complete queries in conversational search. Mo et al. [4] present CHIQ, a two-step method that leverages open-source LLMs to disambiguate conversation history before rewriting, achieving state-of-the-art results on multiple benchmarks without reliance on closed-source models.

2.4 Sequence-Level Knowledge Distillation

Kim and Rush [2] introduce sequence-level knowledge distillation for seq2seq tasks: training a compact student on teacher-generated outputs from beam search, recovering most of the teacher's performance at lower inference cost and often obviating beam search during inference.

3 Method

3.1 Datasets

We experiment on two widely used QA and retrieval benchmarks: MS MARCO Passage Ranking and Natural Questions (NQ).For each clean query, we generate one noisy variant via our noise model (Section 2.2), yielding paired noisy-clean examples for both datasets. All text is lowercased and tokenized using the T5 SentencePiece vocabulary.

3.2 Noise Model

We simulate realistic user errors using the function `noisy_version(query)` (with parameter `p_geom=0.55`). Given a clean query:

1. With probability $1/3$, we leave the query unchanged.
2. Otherwise, we draw

$$X_1 \sim \mathrm{Geom}(p_{\mathrm{geom}}{+}0.1){-}1, \quad X_2 \sim \mathrm{Geom}(p_{\mathrm{geom}}){-}1, \quad X_3 \sim \mathrm{Geom}(p_{\mathrm{geom}}{-}0.1){-}1.$$

3. We then apply, in sequence:
   - **Swap** (X_1 times): pick a random index i and swap characters at positions i and $i + 1$.
   - **Deletion** (X_2 times): delete the character at a uniformly random position.

- **Replacement** (X_3 times): at a random position i, replace the character by a random lowercase neighbor from the simplified QWERTY map (e.g. `{’a’:’qs’, ’b’:’vn’, ..., ’z’:’xs’}`); if no neighbor is defined, use a random letter.

This yields diverse noise patterns such as transpositions, dropouts, and context-aware substitutions. These closely mimic fast typing errors on small keyboards.

3.3 Modeling

Our pipeline consists of two stages:

Teacher fine-tuning We fine-tune a Google Flan-T5-small teacher on the noisy-clean pairs. Training uses standard cross-entropy loss (via the HuggingFace Seq2SeqTrainer), a learning rate of 1×10^{-4}, batch size 8, and 3 epochs. Each input is prefixed with "Rephrase the query by adding synonyms in parentheses for important words (do not repeat):" before feeding into the model.

Sequence-level knowledge distillation We train a Flan-T5-small student to match the teacher's softened logits on the same noisy inputs. At each step, the teacher (frozen) produces logits scaled by $1/\tau$ (with $\tau = 2$), and we minimize the KL divergence between student and teacher distributions, scaled by τ^2. We use a learning rate of 5×10^{-5}, batch size 8, and train for 3 epochs. No explicit cross-entropy term against the clean queries is included.

4 Examples

To illustrate the rewriting quality on realistic noisy queries, Table 1 shows a selection of samples where random character swaps, deletions, and replacements (our noise model) have been applied. For each noisy input, we compare the Flan-T5-base teacher's denoised rewrite against the compact Flan-T5-small student's output. We can see that the student closely matches the teacher in phrasing and inclusion of synonyms, despite its much smaller size.

Across these examples, the student rarely deviates from the teacher's phrasing, showing that distillation successfully transfers nuance. Occasional differences (e.g. slight synonym choices) reflect the student's compact parameter budget but do not harm overall fidelity.

Noisy Query	**Teacher Output**	**Student Output**	**Source**
when did the us become allies with britain	When did the United States form an alliance (partnership, friendship) with the United Kingdom (Great Britain, UK)?	When did the United States become allies with Britain?	natural_questions
are marigolds annuals or perennials	Are marigolds classified as annual (seasonal, one-year) or perennial (long-lived, repeat-blooming) flowers?	Are marigolds annuals or perennials?	ms_marco
what is global migration definition	What is the definition of global migration (international relocation, transnational movement)?	What is the definition of global migration (migration, migrant movement)?	ms_marco
wha is the color of horseshoe crab blood	What is the color of horseshoe crab (marine, species) blood?	What is the color of horseshoe crab blood?	natural_questions
what is teacjing english as a second language	What is the process (method, approach) of teaching English as a Second Language (ESL, foreign language, non-native language)?	What is the meaning of teaching English as a second language (language, dialect)?	natural_questions
how much do providers pay for botox	What is the average reimbursement (payment, compensation) for a botox treatment?	What is the average cost (cost, expense) of a botox treatment?	ms_marco
how long does it take sunlight to reach earth	What is the duration (length, time) it takes for sunlight (solar radiation, sun's rays) to reach the Earth (our planet)?	When does it take sunlight to reach Earth?	ms_marco
what other organs produce amylase	What other organs, aside from the pancreas, produce amylase (diastase, starch-splitting enzyme)?	What other organs produce amylase (adiote, phospholipid)?	ms_marco

Table 1: Sample rewrites for noisy queries: teacher vs. student, with source dataset.

5 Conclusion

We have presented a three-stage pipeline for noise-robust query rewriting that combines a realistic noise simulator, teacher fine-tuning, and sequence-level knowledge distillation. By applying geometric sampling of swap, deletion, and replacement operations, our noise model generates diverse error patterns that closely mimic real user typos. The Flan-T5-base teacher learns high-quality denoising on synthetic noisy-clean pairs, and the distilled Flan-T5-small student recovers most of the teacher's rewriting accuracy while reducing model size and inference latency. Our approach thus makes robust query rewriting practical for deployment in latency-sensitive or resource-constrained environments. Future work includes extending the noise model to capture longer-range errors, evaluating on real user logs, and exploring multilingual setups.

References

[1] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[2] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] Karen Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4), December 1992.

[4] Fang Mo, Bassel Ghaddar, and Others. Chiq: Contextual history enhancement for improving query rewriting in conversational search. In *EMNLP 2024*, 2024.

[5] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek V. Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. *CoRR*, abs/1610.03098, 2016.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[7] Dezhi Ye, Bowen Tian, Jiabin Fan, Jie Liu, Tianhua Zhou, Xiang Chen, Mingming Li, and Jin Ma. Improving query correction using pre-train language model in search engines. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, New York, NY, USA, 2023. Association for Computing Machinery.

[8] Aixin Yu, Chenguang Lin, and Jimmy Lin. Improving multi-turn dialogue modelling with utterance rewriter. In *EMNLP Findings*, 2019.