

University of Windsor

## Scholarship at UWindor

---

Electronic Theses and Dissertations

Theses, Dissertations, and Major Papers

---

9-12-2024

# No Query Left Behind: Query Refinement via Backtranslation

Delaram Rajaei  
*University of Windsor*

Follow this and additional works at: <https://scholar.uwindsor.ca/etd>



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Rajaei, Delaram, "No Query Left Behind: Query Refinement via Backtranslation" (2024). *Electronic Theses and Dissertations*. 9531.

<https://scholar.uwindsor.ca/etd/9531>

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email ([scholarship@uwindsor.ca](mailto:scholarship@uwindsor.ca)) or by telephone at 519-253-3000ext. 3208.

# No Query Left Behind: Query Refinement via Backtranslation

By

**Delaram Rajaei**

A Thesis

Submitted to the Faculty of Graduate Studies  
through the School of Computer Science  
in Partial Fulfillment of the Requirements for  
the Degree of Master of Science  
at the University of Windsor

Windsor, Ontario, Canada

2024

©2024 Delaram Rajaei

No Query Left Behind: Query Refinement via Backtranslation

by

Delaram Rajaei

APPROVED BY:

---

T. Collet-Najem  
Department of Languages, Literatures and Cultures

---

J. Lu  
School of Computer Science

---

H. Fani, Advisor  
School of Computer Science

August 7, 2024

## Declaration of Co-Authorship and Previous Publications

### **I. Co-Authorship**

I hereby declare that this thesis incorporates material that is the result of joint research, as follows:

Chapter 2 of the thesis includes the outcome of publications with the following other co-author, Zahra Taheri under the supervision of Dr. Hossein Fani. Zahra Taheri contributed to the implementation and generation of results for the large language model, along with reviewing the manuscript.

Chapter 3 incorporates submitted material with Zahra Taheri under the supervision of Dr. Hossein Fani. Zahra Taheri contributed by proofreading the manuscript.

Finally, I acknowledge that in all cases the key ideas, primary contributions, experimental designs, data analysis, interpretation, and writing were performed by myself.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

## II. Previous Publication

This thesis includes 2 original papers and 1 poster presentation that have been previously published/submitted for publication in peer-reviewed journals or conferences, as follows:

Thesis Chapter	Publication title/full citation	Publication Status
Chapter 2	Delaram Rajaei, Zahra Taheri, and Hossein Fani, “No Query Left Behind: Query Refinement via Backtranslation”, Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. CIKM '24, October 21–25, 2024, Boise, ID, USA.	Accepted for Publication
Chapter 3	Delaram Rajaei, Zahra Taheri, and Hossein Fani, “Enhancing RAG’s Retrieval via Query Backtranslations”.	Submitted to WISE conference

I certify that I have obtained written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor

## III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone’s copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

## Abstract

Web queries are often brief and unclear due to users’ uncertainty about their information needs, rendering it difficult for search engines to retrieve relevant documents. Query refinement is to enhance the relevance of search results by modifying users’ original queries to *refined* versions. State-of-the-art query refinement models have been trained on web query logs, which are predisposed to topic drifts. To fill the gap, little work has been proposed to generate benchmark datasets of (query  $\rightarrow$  refined query) pairs through an overwhelming application of unsupervised or supervised modifications to the original query while controlling topic drifts. In this paper, however, we propose leveraging natural language backtranslation, a round-trip translation of a query from a source language via target languages, as a simple yet effective *unsupervised* approach to scale up generating gold-standard benchmark datasets. Backtranslation can (1) uncover terms that are omitted in a query for being commonly understood in a source language, but may not be known in a target language (e.g., ‘figs’ $\rightarrow$ (tamil) ‘அத்திமரங்கள்’ $\rightarrow$ ‘the fig [trees](#)’), (2) augment a query with context-aware synonyms in a target language (e.g., ‘italian nobel prize winners’ $\rightarrow$ (farsi) ‘برنده های ایتالیایی جایزه نوبل’ $\rightarrow$ ‘italian nobel [laureates](#)’), and (3) help with the semantic disambiguation of polysemous terms and collocations (e.g., ‘custer’s last stand’  $\rightarrow$ (malay) ‘pertahan terakhir custer’ $\rightarrow$ ‘custer’s last [defence](#)’). Our experiments across 5 query sets with different query lengths and topics and 10 languages from 7 language families using 2 neural machine translators validated the effectiveness of query backtranslation in generating a more extensive gold-standard dataset for query refinement. We open-sourced our research at <https://github.com/fani-lab/RePair/tree/nqlb>.

## Acknowledgement

I am deeply grateful for the unwavering support of my supervisor, whose steadfast belief in my abilities and academic potential encouraged me to persevere even in my lowest moments. His guidance served as a beacon of hope and direction in times of uncertainty. I extend my heartfelt thanks to my amazing teammates, whose collaboration, encouragement, and insights were invaluable throughout this journey. Your dedication and camaraderie transformed challenges into surmountable tasks and made our successes all the more rewarding. This thesis stands as a testament to the unwavering support and boundless love I have received from my family and friends throughout this demanding academic endeavor. My family has always nurtured my curiosity and supported my educational aspirations from the very beginning, and I am immensely grateful for their steadfast support. I am also thankful for my friends, whose timely distractions and unwavering encouragement kept me going when it seemed impossible to continue. In essence, this journey has been a collective effort, enriched by the presence and support of my supervisor, teammates, family, and friends. Their combined influence has not only made this thesis possible but has also made the entire experience fulfilling and memorable.

## Contents

<b>Declaration of Co-Authorship and Previous Publications</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Acknowledgement</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
References . . . . .	7
<b>2 No Query Left Behind: Query Refinement via Backtranslation</b>	<b>13</b>
2.1 Problem Definition . . . . .	13
2.2 Proposed Workflow . . . . .	13
2.2.1 Query Backtranslation . . . . .	14
2.2.2 Query Evaluation . . . . .	15
2.3 Experiments . . . . .	16
2.3.1 Setup . . . . .	18
2.3.1.1 Query Sets. . . . .	18
2.3.1.2 Query Backtranslation. . . . .	18
2.3.1.3 Gold-standard Dataset Generation. . . . .	19
2.3.1.4 Baseline. . . . .	20
2.3.2 Results . . . . .	22
2.3.3 Discussion . . . . .	30
2.4 Concluding Remarks . . . . .	30
References . . . . .	30
<b>3 Enhancing RAG’s Retrieval via Query Backtranslations</b>	<b>37</b>
3.1 Introduction . . . . .	37
3.2 Problem Definition . . . . .	38
3.3 Proposed Approach . . . . .	39
3.3.1 Query Expansion via Natural Language Backtranslation . . . . .	39
3.3.2 RAG-Based Retrieval . . . . .	40
3.4 Experiment . . . . .	40
3.4.1 Dataset . . . . .	41
3.4.2 Baseline . . . . .	42
3.4.3 Setup . . . . .	44
3.4.3.1 Query Backtranslation. . . . .	44
3.4.3.2 RAG-Based Retrieval. . . . .	45



3.4.4	Search and Evaluation . . . . .	46
3.4.5	Results . . . . .	46
3.5	Concluding Remarks . . . . .	51
	References . . . . .	52
<b>4</b>	<b>Poster Presentations</b>	<b>60</b>
4.1	University of Windsor’s 9th Demo Day . . . . .	61
	References . . . . .	61
<b>5</b>	<b>Conclusion</b>	<b>62</b>
5.1	Research Questions . . . . .	62
5.2	Concluding Remarks and Future Work . . . . .	64
	<b>VITA AUCTORIS</b>	<b>66</b>

## List of Tables

1.1.1	Queries and the efficacy of their backtranslations. . . . .	3
2.2.1	Statistics of the query sets; $ q $ shows the length of a query based on the number of terms, $\mathcal{J}$ is the entire set of reference relevant documents (relevance judgments) for queries, and $m_r(q, \mathcal{J}_q) = 1$ indicates queries that need <i>no</i> refinement. . . . .	15
2.3.1	Languages and their families as well as <code>n11b</code> vs. <code>bing</code> 's translation quality; $ q $ shows the length of a query and backtranslation on <code>english</code> is performed for testing the pipeline, which ideally yields the best translation quality. . . . .	17
2.3.2	Results of <code>t5</code> [31] on gold-standard datasets. . . . .	23
2.3.3	Efficacy of backtranslated queries in query refinement. $\#q$ shows the number of original queries that need refinement, while $\#q^*$ and $\%$ represent the <i>best</i> refined queries' count and percentage, respectively, and $\Delta$ denotes the average metric improvements. . . . .	24
2.3.4	Distribution of refined queries across refinement methods, including query backtranslation, local and global unsupervised refiners in terms of <code>map</code> ; $\#q^*$ and $\%$ show the number of best refined queries and percentage, respectively. <b>Bold</b> and underlined numbers are <i>column-wise</i> highest and second-highest among refiners, respectively. . . . .	25
2.3.5	Efficacy of query backtranslation across languages; $\%$ shows the percentage of queries matched with a refined query, and $\Delta$ shows the average metric improvements. <b>Bold</b> and underlined numbers are <i>row-wise</i> highest and second-highest, respectively. . . . .	27
2.3.6	Meta's <code>n11b</code> vs. Microsoft's <code>bing</code> in query refinement. . . . .	29
3.4.1	Query set statistics include query length ( $ q $ ), the full set of relevant documents ( $\mathcal{J}$ ), and queries. . . . .	42

3.4.2 Languages and their families, alongside the translation quality comparison between <b>n11b</b> and <b>bing</b> . Backtranslation into English is tested to ensure optimal translation quality in the pipeline. . . . .	45
3.4.3 <b>rrf</b> vs. non-fused results. . . . .	48
3.4.4 Comparison of the efficacy of rrf-based and original rrf-based results across different datasets. . . . .	49

## List of Figures

1.1.1 Query backtranslation workflow. . . . .	2
2.3.1 The tab-delimited file structure for a gold-standard dataset based on <code>robust04.bm24.map</code> , where <code>-1</code> shows the original query and the rest are refined queries, sorted descending based on the evaluation metric <code>map</code> . . . . .	20
2.3.2 Distribution of $\Delta mrr$ across original queries for <code>backtranslation</code> vs. <code>relevance-feedback</code> , and <code>tagme</code> . . . . .	26
2.3.3 The language spectrum to illustrate the influence of language across each query set based on the number of <i>best</i> refined query obtained by each language. . . . .	28
2.3.4 The length difference between refined query via backtranslation vs. original query. . . . .	28
3.1.1 Generating backtranslated versions of an original query and fusing retrieved document sets for rag-based query refinement. . . . .	39
3.4.1 Effect of different models on refining queries. . . . .	50
3.4.2 Effect of constant <code>k</code> range on fusion outcomes across various categories. . . . .	50
4.1.1 The poster we presented at University of Windsor’s 9th Annual Computer Science Demo Day . . . . .	61

---

# Chapter 1

## *Introduction*

---

### 1.1 Introduction

Retrieving relevant information poses challenges to search engines when user queries are short and unclear, leading to the retrieval of *irrelevant* documents. Query refinement, also known as query expansion or reformulation, aims to transform the user’s original query into a new *refined* version that more accurately reflects the user’s information need and, therefore, improves the relevance of search results. State-of-the-art query refiners are largely based on fine-tuning transformer-based language models [1, 19] or seq-to-seq encoder-decoder neural architecture [2, 9], trained supervisedly on web query logs following *weak* assumptions that users’ input queries improve gradually within a search session, i.e., the last query where the user ends her search session is the refined version of her original query [9]. However, users’ intent may undergo gradual or sudden changes in topics within a search session intrinsically by e.g., search engine’s *incorrect* suggestion of unrelated terms [22], or extrinsically by e.g., online ads, resulting in a loss of sequential semantic context between queries, known as *topic (query) drift* [8, 22]. Also, not all search logs are readily available due to privacy, or when a search engine is newly deployed for a customized application or scarcely used after [6].

Recently, new research efforts have been put into producing gold-standard benchmark datasets that are free of topic drifts and designed specifically to train and evaluate the efficacy of query refiners for web or *non-web* information retrieval systems [26,

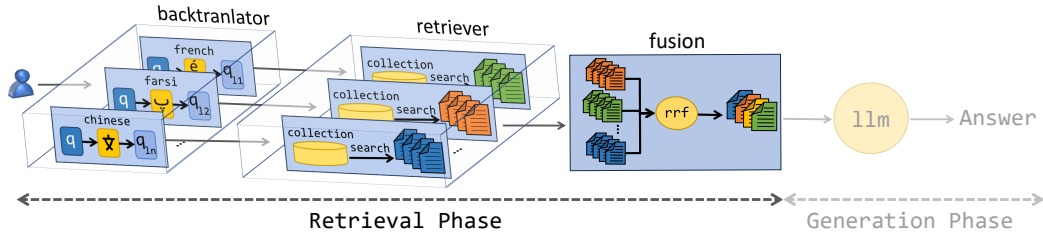


Figure 1.1.1: Query backtranslation workflow.

32, 4]. Tamannaee et al. [26] proposed a pipeline to generate gold-standard datasets from an input set of original queries while controlling topic drift. They applied a host of unsupervised query refiners, from simple lexical lemmatizers to complex pseudo-relevance-based methods, on an original query to generate a wide variety of changes to the query, among which only those that enhance information retrieval metrics like `map` will be chosen as the refined versions of the query. Tamannaee et al.’s pipeline, although comprehensive, rarely finds a refined version; many original queries are left behind with no refined query. Further, it is computationally costly due to the exhaustive application of many refiners on each query. To address scalability, Arabzadeh et al. [4] and others [20] proposed fine-tuning transformer-based language models to generate (query  $\rightarrow$  refined query) pairs. Fine-tuning a transformer, however, demands significant computational resources and time along with its environmental impact [23]. Plus, the efficacy of transformer-based methods is subject to scrutiny given the evaluation data might have been seen during their pre-training, leading to the data leakage threat and a misleading overestimation of their capabilities [28, 15].

In this research, for the first time, we propose to augment such sparse gold-standard datasets even further with more pairs of refined queries using natural language backtranslation; an effective approach that eliminates the need for fine-tuning large transformers and avoids the exhaustive search over many changes to a query. Specifically, we translate an original query from its original language (e.g., **english**) to a target language (e.g., **french**), and then translate it back to the original language using an off-the-shelf neural machine translator (e.g., Meta’s `nllb` [27]) to generate a candidate refined query. While languages share underlying commonalities referred to as linguistic *universals* due to the common neurobiological basis of the human

Table 1.1.1: Queries and the efficacy of their backtranslations.

query id	original query ( $q$ )	(language) translation	backtranslation ( $q_l$ )	map $_{q_l}$ ( $\Delta_{q_l-q}$ )
<b>dbpedia</b>				
SemSearch_ES-13	<i>banana paper making</i>	(korean) 바나나 종이 제조	<i>manufacture of banana paper</i>	1.00 (+0.89)
INEX_XER-116	<i>italian nobel prize winners</i>	(farsi) برنده های ایتالیایی جایزه نوبل	<i>italian nobel laureates</i>	0.57 (+0.34)
INEX_LD-2010057	<i>einstein relativity theory</i>	(swahili) nadharia ya uhusiano wa einstein	<i>einstein theory of relation</i>	0.01 (-0.30)
INEX_LD-20120211	<i>guitar chord tuning</i>	(chinese) 吉他弦調	<i>the guitar strings</i>	0.03 (-0.47)
<b>robust04</b>				
314	<i>marine vegetation</i>	(chinese) 海生植物	<i>the seaweed</i>	0.19 (+0.19)
426	<i>law enforcement, dogs</i>	(swahili) polisi, mbwa	<i>police dogs</i>	0.33 (+0.29)
338	<i>risk of aspirin</i>	(arabic) خطر الأسبرين	<i>the dangers of aspirin</i>	0.15 (-0.25)
403	<i>osteoporosis</i>	(swahili) ugonjwa wa mifupa	<i>bone disease</i>	0.11 (-0.29)
<b>antique</b>				
421753	<i>how to get rid of a skunk?</i>	(swahili) jinsi ya kuondoa skunk?	<i>how to remove skunk</i>	0.25 (+0.05)
1702151	<i>how patient a driver are you?</i>	(french) Vous êtes un chauffeur patient ?	<i>are you a patient driver?</i>	0.35 (+0.12)
204633	<i>why do you have memories?</i>	(korean) 왜 기억이 나나요?	<i>why do you remember</i>	0.00 (-0.11)
1944018	<i>why cannot teenagers vote?</i>	(tamil) ஏன் இளைஞர்கள் வாக்களிக்க முடியாது?	<i>why cannot young people vote</i>	0.04 (-0.10)
<b>gov2</b>				
804	<i>ban on human cloning</i>	(farsi) ممنوعیت کlon کردن انسان	<i>prohibition of human cloning</i>	1.00 (+0.48)
822	<i>custer's last stand</i>	(malay) pertahanan terakhir custer	<i>custer's last defense</i>	0.13 (+0.03)
753	<i>bullying prevention programs</i>	(french) programmes de prévention de l'intimidation	<i>the prevention of bullying programmes</i>	0.06 (-0.03)
757	<i>murals</i>	(german) wandmalereien	<i>wall paintings</i>	0.00 (-0.04)
<b>clueweb09b</b>				
154	<i>figs</i>	(tamil) அத்திமரங்கள்	<i>the fig trees</i>	1.00 (+0.91)
130	<i>fact on uranus</i>	(korean) 천왕성에 대한 사실	<i>the facts about uranus</i>	0.16 (+0.01)
51	<i>horse hooves</i>	(farsi) کفش اسب	<i>horse shoes</i>	0.03 (-0.19)
75	<i>tornadoes</i>	(arabic) الأعاصير	<i>hurricanes</i>	0.00 (-0.08)

brain [12], they carry differences on the surface, including phonetics, morphological units (terms), syntax, and semantics to convey pragmatics and establish a discourse, especially in an informal context like in ad-hoc web queries, that can be leveraged via backtranslation to generate diverse paraphrases of a query while withholding semantic [31]:

- Backtranslation can uncover terms or entities that are latent in a query for being superfluous or part of background knowledge in a source language, also known as ellipsis [7]. However, such latent terms may *not* be commonly known in a target language, and hence, they should be explicitly generated through translation. For instance, from Table 1.1.1, when the short query ‘figs’ is translated to **tamil** as ‘அத்திமரங்கள்’ followed by a backtranslation to **english** as ‘the fig trees’, it brings up ‘trees’ and enhanced **bm25**’s **map** from 0.04 to 0.07;

- Backtranslation can effectively augment *context-aware* synonymous terms from a target language to the original query, as opposed to simple synonym replacement by a traditional query refiner [25]. For instance, when *italian nobel prize winners* is translated to `farsi` as `برنده های ایتالیایی جایزه نوبل`, followed by a backtranslation to `english` as *italian nobel laureates*, it brings up *laureates* for *prize winners* as opposed to *medalist* or *champions*, which increased the `map` for `bm25` from 0.22 to 0.56;
- Backtranslation can disambiguate polysemous terms and collocations. For instance, translating *custers last stand*<sup>1</sup> to `malay` *pertahan terakhir custer*, and backtranslating to `english`, *custers last defence* maps the term *stand* to *defence*, which is more semantically related to the wars and battles, leading to the detection of the latent context of a *battle* and a `map` improvement from 0.10 to 0.13, as opposed to other semantics like *political stand* or *upright body position*;

For similar reasons, backtranslation has been employed in review analysis and opinion mining [11, 30, 18, 14] and other natural language processing tasks like text summarization [10] and question-answering [5], and machine translation [13, 24, 17]. Furthermore, the open-source accessibility to multilingual neural machine translators [29, 16, 27], capable of delivering high-quality translations between many languages, including low-resource languages, as well as their smooth integration into any pipeline with few lines of code, have already set off a surge of interests in backtranslation.

In this research, we proposed a reproducible domain-agnostic pipeline to generate refined queries via language backtranslation. From Figure 3.1.1, our pipeline takes as input: (1) a query set in a source language, e.g., `english` along with relevance judgments, (2) a set of target languages, e.g., `{farsi, chinese, ...}`, (3) an information retrieval method, e.g., `bm25` and (4) an evaluation metric (e.g., `map`), and outputs a golden dataset that includes pairs of  $(q \rightarrow q^*)$  such that  $q^*$  retrieves better search results compared to  $q$  under the information retrieval method and the evaluation metric. Our findings show that query backtranslation substantially expands

---

<sup>1</sup>[https://en.wikipedia.org/wiki/battle\\_of\\_the\\_little\\_bighorn](https://en.wikipedia.org/wiki/battle_of_the_little_bighorn)



gold-standard datasets for supervised query refinement while outperforming existing unsupervised refiners across query sets from various domains with different query lengths and diverse topics. The efficacy of the expanded datasets with query back-translations has further been evidenced via the performance boost of a fine-tuned large language model (`t5` [21]). Our findings also underline the choice of a translator; a translator may fall short of query refinement should it translate accurately but with little to no diversity in generating new query terms during query backtranslation.

In summary, our main findings show that:

1. Backtranslation of languages can significantly enhance the creation of gold-standard datasets for query refinement, as demonstrated by its effectiveness in improving evaluation metrics. On average, the scalability results per query set are as follows: 47.62% improvement in `dbpedia`, 48.51% in `robust04`, 39.55% in `antique`, 36.06% in `gov2`, and 17.62% in `clueweb09b`.
2. Backtranslation outperforms unsupervised refiners, by scaling up the number of refined queries compared to the 22 unsupervised refiners by Tamannaee et al. [26] across different domains and metrics.
3. The effectiveness of backtranslation remains consistent across languages originating from diverse language families. Our research reveals that the semantic coherence of backtranslated queries is shaped by the linguistic correlation between the source and target languages. We find that using languages from the same family such as `french` yields more semantically related queries, whereas employing languages from different families such as `chinese` generates a wider range of diverse outputs.
4. The effectiveness of backtranslation remains consistent across query sets from various domains. We attribute the domain-specific performance of languages in query refinement to i) the size of queries affecting the quality of backtranslation, and ii) the diversity of topics within query sets. Our findings suggest that while backtranslation is effective for refining diverse queries its efficacy may diminish when

applied to short queries derived from a corpus with a large range of topics, such as `clueweb09b` compared to `dbpedia` and `robust04`.

5. The effectiveness of backtranslation in query refinement is closely linked to the selection of a neural machine translator. Our findings indicate that the choice of translator significantly influences the outcomes of the refinement process. A more accurate translator tends to yield superior translations, resulting in a greater number of refined queries.

In summary, our main contributions lie in the following:

1. We propose natural language backtranslation augmentation for query refinement. We show that query backtranslation not only effortlessly expands gold-standard datasets for training supervised query refinement methods but also is a strong unsupervised method for query refinement;
2. We study query backtranslation across diverse languages from different language families<sup>2</sup>, including `french`, `german`, `russian`, and `farsi` from indo-european, `malay` from austronesian, `tamil` from dravidian, `swahili` from bantu, `chinese` from sino-tibetan, `korean` from koreanic, and `arabic` from afro-asiatic;
3. We benchmark query backtranslation across five prominent `trec` query sets spanning diverse domains, including `dbpedia` for wikipedia articles, `robust04` for news articles, `antique` for yahoo’s question-answering community, and `gov2` and `clueweb09b` for web queries.
4. We fine-tune `t5` [21], a well-known unified language model for transfer learning in nlp tasks, on the datasets expanded by query backtranslations, and lack thereof, for the task of supervised query refinement. We show that the expanded datasets effectively improve the model’s performance in predicting refined queries in terms of information retrieval metrics.

---

<sup>2</sup>A language family is a set of languages which share cultural roots and exhibit similarities in vocabulary and grammar [3].

5. We open-sourced our pipeline to support the reproducibility of our research. By making our pipeline openly accessible, we encourage collaboration and enable others to validate and build upon our findings.

## References

- [1] Sanjay Agrawal, Srujana Merugu, and Vivek Sembium. “Enhancing E-commerce Product Search through Reinforcement Learning-Powered Query Reformulation”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. Ed. by Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos. ACM, 2023, pp. 4488–4494. DOI: 10.1145/3583780.3615474. URL: <https://doi.org/10.1145/3583780.3615474>.
- [2] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. “Context attentive document ranking and query suggestion”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in IR*. 2019, pp. 385–394.
- [3] Stephen R. Anderson. “10How many languages are there in the world?” In: *Languages: A Very Short Introduction*. 2012. ISBN: 9780199590599.
- [4] Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. “Matches Made in Heaven: Toolkit and Large-Scale Datasets for Supervised Query Reformulation”. In: 2021, pp. 4417–4425.
- [5] Shabbirhussain Bhaisaheb, Shubham Paliwal, Rajaswa Patil, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. “Program Synthesis for Complex QA on Charts via Probabilistic Grammar Based Filtered Iterative Back-Translation”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2501–2515. DOI: 10.18653/

- v1/2023.findings-eacl.189. URL: <https://aclanthology.org/2023.findings-eacl.189>.
- [6] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. “Query suggestions in the absence of query logs”. In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. Ed. by Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft. ACM, 2011, pp. 795–804. DOI: 10.1145/2009916.2010023. URL: <https://doi.org/10.1145/2009916.2010023>.
- [7] Damir Cavar, Ludovic Mompelat, and Muhammad Abdo. “The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications”. In: *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*. Ed. by Michael Hahn et al. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 46–54. URL: <https://aclanthology.org/2024.sigtyp-1.6>.
- [8] Jia Chen, Jiabin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. “Towards a Better Understanding of Query Reformulation Behavior in Web Search”. In: *WWW ’21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. Ed. by Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia. ACM / IW3C2, 2021, pp. 743–755. DOI: 10.1145/3442381.3450127. URL: <https://doi.org/10.1145/3442381.3450127>.
- [9] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. “Learning to attend, copy, and generate for session-based query suggestion”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 1747–1756.
- [10] Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq R. Joty, Dragomir R. Radev, and Yashar Mehdad. “Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation”. In: *Proceedings of the 2021 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Association for Computational Linguistics, 2021, pp. 704–717. DOI: 10.18653/v1/2021.naacl-main.57. URL: <https://doi.org/10.18653/v1/2021.naacl-main.57>.
- [11] Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. “Latent Target-Opinion as Prior for Document-Level Sentiment Classification: A Variational Approach from Fine-Grained Perspective”. In: *WWW ’21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. Ed. by Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia. ACM / IW3C2, 2021, pp. 553–564. DOI: 10.1145/3442381.3449789. URL: <https://doi.org/10.1145/3442381.3449789>.
- [12] Angela D Friederici. *Language in our brain: The origins of a uniquely human capacity*. MIT Press, 2017.
- [13] Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. “Revisiting Iterative Back-Translation from the Perspective of Compositional Generalization”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 7601–7609. DOI: 10.1609/AAAI.V35I9.16930. URL: <https://doi.org/10.1609/aaai.v35i9.16930>.
- [14] Farinam Hemmatizadeh, Chrsitine Wong, Alice Yu, and Hossein Fani. “Latent Aspect Detection via Backtranslation Augmentation”. In: *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management, University of Birmingham and Eastside Rooms, UK, October 21-25, 2023*. ACM, 2023. DOI: 10.1145/3583780.3615205.

- [15] Xing Hu, Ling Liang, Xiaobing Chen, Lei Deng, Yu Ji, Yufei Ding, Zidong Du, Qi Guo, Timothy Sherwood, and Yuan Xie. “A Systematic View of Model Leakage Risks in Deep Neural Network Systems”. In: *IEEE Trans. Computers* 71.12 (2022), pp. 3254–3267. DOI: 10.1109/TC.2022.3148235. URL: <https://doi.org/10.1109/TC.2022.3148235>.
- [16] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 67–72. URL: <https://www.aclweb.org/anthology/P17-4012>.
- [17] Yu Li, Xiao Li, Yating Yang, and Rui Dong. “A Diverse Data Augmentation Strategy for Low-Resource Neural Machine Translation”. In: *Inf.* 11.5 (2020), p. 255. DOI: 10.3390/info11050255. URL: <https://doi.org/10.3390/info11050255>.
- [18] Tomas Liesting, Flavius Frasinca, and Maria Mihaela Truşcă. “Data Augmentation in a Hybrid Approach for Aspect-Based Sentiment Analysis”. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC ’21. Virtual Event, Republic of Korea: Association for Computing Machinery, 2021, pp. 828–835. ISBN: 9781450381048. DOI: 10.1145/3412841.3441958. URL: <https://doi.org/10.1145/3412841.3441958>.
- [19] Akash Kumar Mohankumar, Nikit Begwani, and Amit Singh. “Diversity driven Query Rewriting in Search Advertising”. In: *KDD ’21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*. Ed. by Feida Zhu, Beng Chin Ooi, and Chunyan Miao. ACM, 2021, pp. 3423–3431. DOI: 10.1145/3447548.3467202. URL: <https://doi.org/10.1145/3447548.3467202>.
- [20] Yogeswar Lakshmi Narayanan and Hossein Fani. “RePair: An Extensible Toolkit to Generate Large-Scale Datasets via Transformers for Query Refinement”. In: *Proceedings of the 32nd ACM International Conference on Information &*

- Knowledge Management, University of Birmingham and Eastside Rooms, UK, October 21-25, 2023*. ACM, 2023. DOI: 10.1145/3583780.3615129.
- [21] Colin Raffel and et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.
- [22] Haggai Roitman, Ella Rabinovich, and Oren Sar Shalom. “As Stable As You Are: Re-ranking Search Results using Query-Drift Analysis”. In: *Proceedings of the 29th on Hypertext and Social Media, HT 2018, Baltimore, MD, USA, July 09-12, 2018*. Ed. by Dongwon Lee, Nishanth Sastry, and Ingmar Weber. ACM, 2018, pp. 33–37. DOI: 10.1145/3209542.3209567. URL: <https://doi.org/10.1145/3209542.3209567>.
- [23] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. “Green ai”. In: *Communications of the ACM* 63.12 (2020), pp. 54–63.
- [24] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Improving Neural Machine Translation Models with Monolingual Data”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. DOI: 10.18653/v1/p16-1009. URL: <https://doi.org/10.18653/v1/p16-1009>.
- [25] Ali Asghar Shiri. “End-user interaction with thesaurus-enhanced search interfaces, an evaluation of search term selection for query expansion”. In: (2003).
- [26] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. “Reque: a configurable workflow and dataset collection for query refinement”. In: 2020, pp. 3165–3172.
- [27] NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. “No language left behind: Scaling human-centered machine translation”. In: *línea*. Disponible en: <https://github.com/facebookresearch/fairseq/tree/nllb> (2022).

- [28] Yonghao Wu, Zheng Li, Jie M. Zhang, and Yong Liu. “ConDefects: A New Dataset to Address the Data Leakage Concern for LLM-based Fault Localization and Program Repair”. In: *CoRR* abs/2310.16253 (2023). DOI: 10.48550/ARXIV.2310.16253. arXiv: 2310.16253. URL: <https://doi.org/10.48550/arXiv.2310.16253>.
- [29] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). arXiv: 1609.08144. URL: <http://arxiv.org/abs/1609.08144>.
- [30] Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. “Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 9298–9305. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6469>.
- [31] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. “Qanet: Combining local convolution with global self-attention for reading comprehension”. In: *arXiv preprint arXiv:1804.09541* (2018).
- [32] George Zerveas, Ruochen Zhang, Leila Kim, and Carsten Eickhoff. “Brown University at TREC Deep Learning 2019”. In: *CoRR* abs/2009.04016 (2020). arXiv: 2009.04016. URL: <https://arxiv.org/abs/2009.04016>.



---

# Chapter 2

## *No Query Left Behind: Query*

### *Refinement via Backtranslation*

DELARAM RAJAEI, ZAHRA TAHERI, HOSSEIN FANI

Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA

---

#### 2.1 Problem Definition

Given an original query  $q$  along with its reference set of relevant documents (relevance judgment)  $\mathcal{J}_q$ , an information retrieval method (retriever)  $r$ , and an evaluation metric  $m$ , which measures the quality of  $r$  for the query  $q$ , denoted by  $m_r(q, \mathcal{J}_q) \in \mathbb{R}^{[0,1]}$ , and  $m_r(q, \mathcal{J}_q) < 1$ , query refinement aims at identifying the set of *refined* versions  $\mathcal{R}_{q,r,m} = \{q^\diamond\}$  for  $q$  such that  $m_r(q, \mathcal{J}_q) < m_r(q^\diamond, \mathcal{J}_q); \forall q^\diamond \in \mathcal{R}_{q,r,m}$ , that is,  $q^\diamond$  retrieve more relevant documents under  $r$  in terms of  $m$ . We also denote  $q^*$  to the *best* refined query, i.e.,  $q^* = \operatorname{argmax}_{q^\diamond \in \mathcal{R}_{q,r,m}} m_r(q^\diamond, \mathcal{J}_q)$ . We refer to  $q$  as a *hard* query, denoted by  $\bar{q}$ , when query refinement falls short of finding any refined version, i.e.,  $\mathcal{R}_{\bar{q},r,m} = \emptyset$ . An original query  $q$  might be the best query in the first place, i.e.,  $m_r(q, \mathcal{J}_q) = 1$  and  $\mathcal{R}_{q,r,m} = q = q^*$ , and hence, query refinement is unnecessary.

#### 2.2 Proposed Workflow

In this section, we describe our proposed configurable workflow to scale up the generation of gold-standard datasets for the supervised query refinement task via our novel application of natural language backtranslation. The overview of our proposed work-

flow is shown in Figure 3.1.1. The input of our workflow is a set of original unrefined queries and their associated relevance judgements, as well as an information retrieval method or a retriever, e.g., `bm25` and an evaluation metric, e.g., `map`. The output of this process is a ranked list of refined queries for each original query, each of which effectively improves the performance of the information retrieval method in terms of the given evaluation metric. The proposed workflow includes two main components: (1) query backtranslation and (2) query evaluation, detailed hereafter.

### 2.2.1 Query Backtranslation

Natural languages are the primary vehicle for communication, allowing thoughts to be efficiently shared between humans, conveying the culture, history, and heritage of a common people [9, 14]. While languages share underlying commonalities referred to as linguistic *universals* due to the common neurobiological basis of the human brain [11], they carry differences on the surface to convey similar pragmatics and discourse, especially in an informal context. Prominent examples are gendered pronouns, phrases, proverbs, and particularly *ellipses* in writing when we omit terms or phrases that are nevertheless understood in the context of the remaining terms or common background knowledge [5]. In query backtranslation, we aim to benefit from languages’ differences on the surface while conveying the same or similar underlying semantics for a query in a source language via a round-trip translation to a target language (forward translation) and translating the result back into the source language (backward translation). We presume that backtranslation preserves the query’s semantic context, yet (1) can uncover latent occurrences of entities (ellipses) because a latent entity may not be part of background knowledge in a target language and will be explicitly generated through backtranslation, which can be kept after the backtranslation to the original query, (2) augments context-aware synonyms to the original query from a target language, and (3) helps with the semantic disambiguation of polysemous terms and collocations. As shown in Table 1.1.1, a backtranslated version of a query may carry term replacement (e.g., ‘[manufacture](#) of banana paper’ for ‘banana paper making’ in backtranslation through `korean`) where the term ‘mak-

Table 2.2.1: Statistics of the query sets;  $|q|$  shows the length of a query based on the number of terms,  $\mathcal{J}$  is the entire set of reference relevant documents (relevance judgments) for queries, and  $m_r(q, \mathcal{J}_q) = 1$  indicates queries that need *no* refinement.

		avg $m_r(q, \mathcal{J}_q)$						$m_r(q, \mathcal{J}_q) = 1$							
		bm25			qld			bm25			qld				
query set	domain	#q	#documents	avg  q	\mathcal{J}_q	#q : \mathcal{J}_q =	map	ndcg	mrr	map	ndcg	mrr	map	ndcg	mrr
dbpedia [1, 16]	wikipedia	467	4,635,922	5.37	49,280	0	0.232	0.392	0.565	0.292	0.469	0.663	7	6	212
robust04 [42]	news	250	528,155	2.76	311,410	1	0.199	0.368	0.667	0.201	0.373	0.681	1	1	138
antique [15]	non-factoid questions	200	403,666	9.34	6,589	0	0.353	0.494	0.881	0.252	0.420	0.729	0	0	163
gov2 [7]	*.gov web	150	1,247,753	3.13	135,352	1	0.157	0.317	0.718	0.165	0.324	0.706	1	1	93
clueweb09b [6]	web	200	50,000,000	2.45	84,366	2	0.078	0.180	0.383	0.073	0.172	0.304	2	2	55

*ing*’ is replaced by *manufacture*) and/or new terms, (e.g., ‘*figs*’ is expanded with the term ‘*trees*’ in ‘*the fig [trees](#)*’ in backtranslation through `tamil`), which yield more effective information retrieval.

Formally, let  $\mathcal{L}$  be the set of natural languages. Given an original query  $q$  in a source language, we translate it to a target language  $l \in \mathcal{L}$  and backtranslate the result to the source language, which results in a backtranslated and possibly modified version of the query, denoted by  $q_l$ , which may or may *not* be a refined query. We generate the set of backtranslated versions of the  $q$  via all languages  $\mathcal{L}$  languages  $q_{\mathcal{L}} = \{q_l : \forall l \in \mathcal{L}\}$ .

To perform forward and backward query translations, we utilize a neural machine translator that (1) is capable of providing high-quality *two-way* translations between a wide variety of languages, including low-resource ones, to enable comprehensive study on query backtranslation via languages with distinct properties, (2) is open-sourced to foster transparency, and (3) can be smoothly integrated into our pipeline with few lines of code. Examples include Meta’s ‘*no language left behind*’ (`nllb`) [40], an open-source neural machine translator between two hundred languages with a particular focus on realizing a universal translation system while prioritizing low-resource languages, as opposed to a small dominant subset of languages.

## 2.2.2 Query Evaluation

Given an original query  $q$ , we evaluate the backtranslated queries to select the *refined* ones as the improved queries. Given the relevance judgment  $\mathcal{J}_q$ , a backtranslated

query  $q_l$  is evaluated based on how it improves the performance of the given information retrieval method  $r$  with respect to an evaluation metric  $m$  and will be selected as a refined query  $q^\diamond$  for the set  $\mathcal{R}_{q,r,m}$ . Formally:

$$\mathcal{R}_{q,r,m} = \{q^\diamond \quad : \quad q_l \in q_{\mathcal{L}}, m_r(q, \mathcal{J}_q) < m_r(q_l, \mathcal{J}_q)\} \quad (1)$$

where  $m_r(\cdot, \mathcal{J}_q)$  is the performance of the information retrieval method  $r$  over a query, measured by the evaluation metric  $m$ , and with respect to the relevance judgments for query  $q$ . Simply put, the elements in  $\mathcal{R}_{q,r,m}$  are those queries  $q_l \in q_{\mathcal{L}}$  for which retrieval method  $r$  has retrieved better results in comparison to the results it has retrieved using the original unrefined query  $q$ .

## 2.3 Experiments

In this section, we present the details of our experiments toward addressing the following research questions:

**RQ1: Can language backtranslation *effectively* scale up generating gold-standard datasets for query refinement?** We implement backtranslation via 10 languages across 7 language families, including low-resource languages, as refinement techniques within our pipeline to answer this question. We evaluate the performance of the backtranslated queries using 2 information retrieval methods and 3 evaluation metrics. To assess the efficacy of backtranslation for query refinement, we calculate how many of the backtranslated queries become refined queries as well as to what extent they improve each evaluation metric. To show whether the scale-up is indeed effective for supervised methods, we fine-tuned a large language model using the generated datasets with backtranslations and lack thereof.

**RQ2: How does backtranslation fare vs. unsupervised refiners?** We compared refined queries resulting from backtranslation against 22 unsupervised refiners across different information retrieval methods, evaluation metrics and query sets from various domains.

2. NO QUERY LEFT BEHIND: QUERY REFINEMENT VIA BACKTRANSLATION

Table 2.3.1: Languages and their families as well as nllb vs. bing’s translation quality;  $|q|$  shows the length of a query and backtranslation on english is performed for testing the pipeline, which ideally yields the best translation quality.

family	language	dbpedia						robust04						antique						gov2						clueweb09b					
		$ q  -  q $		declutr [12]		rouge-1		$ q  -  q $		declutr [12]		rouge-1		$ q  -  q $		declutr [12]		rouge-1		$ q  -  q $		declutr [12]		rouge-1		$ q  -  q $		declutr [12]		rouge-1	
		nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing
	english	+0.01	+0.01	1.00	1.00	1.00	1.00	-0.11	-0.11	1.00	1.00	1.00	1.00	-0.10	-0.10	1.00	1.00	1.00	1.00	-0.07	-0.07	1.00	1.00	1.00	1.00	+0.01	+0.01	1.00	1.00	1.00	1.00
	farsi	+0.54	+0.09	0.83	0.85	0.62	0.75	+0.77	+0.09	0.81	0.85	0.52	0.72	-0.41	-0.36	0.84	0.86	0.63	0.76	+1.02	+0.24	0.79	0.86	0.47	0.70	+0.76	+0.01	0.74	0.80	0.54	0.73
indo-european	french	+0.37	+0.16	0.87	0.86	0.70	0.81	+0.91	+0.35	0.85	0.86	0.56	0.75	-0.14	0.00	0.89	0.89	0.72	0.81	+1.02	+0.41	0.82	0.87	0.52	0.75	+0.48	+0.11	0.81	0.83	0.60	0.84
	german	+0.63	+0.11	0.85	0.87	0.72	0.83	+1.06	+0.39	0.81	0.86	0.54	0.74	-0.28	+0.20	0.89	0.89	0.73	0.82	+1.13	+0.47	0.79	0.87	0.53	0.73	+0.85	+0.19	0.75	0.83	0.59	0.83
	russian	+0.43	+0.21	0.86	0.86	0.69	0.79	+0.79	+0.42	0.84	0.85	0.56	0.70	-0.36	-0.09	0.88	0.86	0.69	0.78	+1.14	+0.46	0.81	0.86	0.49	0.68	+0.62	+0.09	0.77	0.82	0.54	0.79
austronesian	malay	+0.26	+0.08	0.88	0.88	0.69	0.77	+0.48	+0.14	0.85	0.88	0.57	0.70	-0.09	-0.16	0.88	0.90	0.70	0.81	+0.74	+0.25	0.85	0.90	0.53	0.70	+0.36	+0.03	0.82	0.84	0.63	0.80
dravidian	tamil	+1.64	+0.03	0.84	0.86	0.62	0.81	+1.20	+0.06	0.81	0.87	0.50	0.75	-0.16	+0.27	0.86	0.87	0.64	0.76	+0.88	+0.18	0.82	0.88	0.49	0.79	+0.69	+0.04	0.77	0.82	0.56	0.85
bantu	swahili	+0.21	0.00	0.87	0.87	0.69	0.77	+0.69	+0.23	0.82	0.86	0.49	0.67	-0.28	-0.07	0.88	0.87	0.68	0.76	+1.02	+0.23	0.79	0.90	0.44	0.76	+0.38	+0.04	0.81	0.84	0.59	0.80
sino-tibetan	chinese	+1.75	+0.20	0.80	0.86	0.51	0.71	+0.95	+0.26	0.78	0.87	0.45	0.69	-1.02	-0.04	0.84	0.86	0.59	0.73	+0.95	+0.34	0.77	0.87	0.43	0.64	+0.82	+0.17	0.72	0.82	0.42	0.70
koreanic	korean	+0.53	+0.14	0.82	0.85	0.58	0.73	+1.36	+0.17	0.80	0.84	0.47	0.70	+1.07	-0.13	0.83	0.87	0.59	0.75	+1.03	+0.21	0.78	0.86	0.43	0.68	+1.01	+0.22	0.74	0.81	0.53	0.74
afro-asiatic	arabic	+0.42	+0.06	0.83	0.87	0.65	0.77	+2.36	+0.24	0.78	0.86	0.53	0.74	-0.11	-0.23	0.86	0.87	0.68	0.79	+0.94	+0.29	0.77	0.87	0.46	0.69	+0.78	-0.02	0.72	0.83	0.51	0.82

**RQ3: Is the efficacy of backtranslation consistent across languages from different language families?** We perform a comparative analysis on languages from 7 families. Our objective is to study whether the semantic coherence of the backtranslated queries is influenced by the linguistic relationship between the source and target languages. We expect more semantically related queries if the source and target languages are in the same family. Conversely, we hypothesize that utilizing source and target languages from different language families may result in the generation of more diverse outputs. By comparing the outcomes across these languages, we aim to uncover any visible patterns or variations in the efficacy of backtranslation. This analysis provides valuable insights into the cross-linguistic performance of backtranslation.

**RQ4: Is the efficacy of backtranslation consistent across query sets from different domains?** As for this question, we generate query backtranslations for 5 query sets withholding various query lengths, short vs. long queries, and topics in different domains, news articles vs. web.

**RQ5: Does the efficacy of query backtranslation depend on the choice of a neural machine translator?** To address this inquiry, we conduct experiments across two neural machine translators, which are built on different technologies and platforms, namely nllb [40] and bing [24].

### 2.3.1 Setup

#### 2.3.1.1 Query Sets.

Our benchmark includes well-known query sets in `english` from different domains, namely `dbpedia` [1, 16] collection of wikipedia articles, `robust04` [42] collection of news articles and US government publications, `antique`'s test collection [15] including open-domain non-factoid questions from Yahoo! Answers, `gov2` [7] webpages of `.gov` web domain, and `clueweb09b` [6] collection of webpages. In all query sets, we filter out queries with *no* relevance judgment. Also, given an information retrieval method and an evaluation metric, we skip those queries that result in the best metric value of 1.00, for no refinement is needed. Table 3.4.1 summarizes the statistics of the query sets. As seen in `robust04`, `gov2`, and `clueweb09b`, the average query lengths are 2.76, 3.13, and 2.45, respectively, indicating relatively short queries. Conversely, `antique` exhibits longer queries, with an average length of 9.34 terms, suggesting more detailed or complex information needs, and `dbpedia` falls within an intermediate range with average query lengths of 5.37 terms.

#### 2.3.1.2 Query Backtranslation.

We leverage Meta's '*no language left behind*' (`nllb`) [40]<sup>1</sup>, for being open-source, capable of providing two-way translations in 200 languages with a focus on low-resource languages, and easily integrated into any pipeline with few lines of code. Meta's `nllb` is available with model card [26] and is developed based on a conditional mixture of several transformers [35] that is trained on data tailored for low-resource languages. On the other extreme, we alternatively chose the `bing` translator<sup>2</sup>, a cloud-based *closed*-source machine translation service offered by Microsoft [23, 24] which supports around 128 languages, yet has *no* publicly available model card and/or documentation, to the best of our search. We deliberately aim to compare the efficacy of our method via two extremes of a well-documented translator against a relatively opaque/obscure translator.

<sup>1</sup> <https://github.com/facebookresearch/fairseq/tree/nllb>

<sup>2</sup><https://www.bing.com/Translator>

We translate queries from `english` into 10 languages from 7 language families, including `malay`, `swahili`, and `tamil` as low-resource languages. Table 3.4.2 shows the average difference between the number of terms in the original queries in `english` and the backtranslated versions via different languages ( $|q_l| - |q|$ ) as well as the average pairwise similarities between a query and its backtranslated versions using `rouge-1` [21] and `declutr` by Giorgi et al. [12]. Backtranslation from `english` to itself has been performed for unit test purposes where all the results for `declutr` and `rouge-1` are expected to be the highest possible 1.0 with a negligible change in query length. As seen, all languages could expand the original queries of query sets with new terms in the backtranslated versions with an exception in `antique` set where queries are long questions and backtranslation versions are of the same or contracted lengths, while the semantics remained almost surely intact in terms of `rouge-1` and `declutr` scores. In terms of translation quality, while `rouge-1` considers the overlap of n-grams between a pair of an original and backtranslated query, and hence, falls short of capturing topic drifts, if any, `declutr` relies on the cosine similarity between a pair of query embeddings in a *latent space* and is more effective in measuring semantic similarities. Comparing `nllb` and `bing`, while both translators obtain similar performance in terms of the `declutr`, `bing` has higher values of `rouge-1` indicating *fewer* new terms and *less* diverse paraphrases in backtranslated queries, which yield its poorer performance for query refinement task, as will be discussed when answering **RQ5**.

### 2.3.1.3 Gold-standard Dataset Generation.

We have applied two *sparse* information retrieval methods, namely `bm25` [32] and `qld` [30], using `pyserini` [22] to retrieve relevant content for the original queries as well as the backtranslated versions. We acknowledge dense information retrieval methods like `colbert` [17] and their state-of-the-art retrieval performance. However, we intentionally exclude them in this paper due to their extreme time, space, and computation resource consumption to vectorize an entire collection of documents in our query sets. Further, herein, our main goal is to show the novel application of backtranslation in scaling up the gold-standard datasets for supervised query refinement

qid	order	query	bm25.map
304	-1	endangered species (mammals)	0.0591
304	bt_nllb_swahili	endangered species animals	0.0698
304	bt_nllb_korean	endangered species	0.0624
304	bt_nllb_farsi	endangered species clover	0.0600

Figure 2.3.1: The tab-delimited file structure for a gold-standard dataset based on `robust04.bm24.map`, where `-1` shows the original query and the rest are refined queries, sorted descending based on the evaluation metric `map`.

methods, which can be achieved even with off-the-shelf lightweight retrievers; with better dense retrievals, better efficacy in query backtranslation would be expected. That said, we will obtain the results for dense retrieval in the future to enrich our findings further.

We evaluate the retrieval performances based on three metrics, i.e., `map`, `mrr` and `ndcg`, using `trec_eval` [29]. Those backtranslated versions that increased a metric value form a gold-standard dataset. In total, we generate a family of  $\{\text{dbpedia}, \text{robust04}, \text{antique}, \text{gov2}, \text{clueweb09b}\} \times \{\text{bm25}, \text{qld}\} \times \{\text{map}, \text{mrr}, \text{ndcg}\} = 30$  gold-standard datasets. Figure 2.3.1 shows the file structure of the gold-standard dataset in `robust04.bm25.map.tsv`.

#### 2.3.1.4 Baseline.

To demonstrate the efficacy of query backtranslation, we present two sets of comparative baselines. (1) We compare our backtranslation pipeline with *global* and *local* unsupervised refinement methods in generating gold-standard datasets for training supervised or semi-supervised query refinement methods. It is worth noting that supervised query refinement methods cannot be a baseline herein as they rely on the training datasets that we aim to generate via unsupervised methods.

Global methods consider an original query only, and include:

- `tagme` [10], which replace the original query’s terms with the title of their `wikipedia` articles,
- stemmers, which utilize various lexical, syntactic, and semantic aspects of query terms and their relationships to reduce the terms to their roots, including `krovetz`,



lovins, paiceHusk, porter, sremoval, trunc4, and trunc5 [34],

- semantic refiners, which use an external linguistic knowledge-base including `thesaurus` [37], `wordnet` [28], and `conceptnet` [8], to extract related terms to the original query’s terms,
- `sense-disambiguation` [39], which resolves the ambiguity of polysemous terms in the original query based on the surrounding terms and then adds the synonyms of the query terms as the related terms,
- embedding-based methods, which use pre-trained term embeddings from `glove` [13] and `word2vec` [25] to find the most similar terms to the query terms,
- `anchor` [18], which is similar to embedding methods where the embeddings trained on wikipedia articles’ *anchors*, presuming an anchor is a concise summary of the content in the linked page,
- `wiki` [36], which uses the embeddings trained on wikipedia’s hierarchical categories [20] to add the most similar concepts to each query term.

Local refiners, however, consider terms from top- $k$  retrieved documents via a prior retrieval using an information retrieval method, e.g., `bm25` or `qld`, to find an initial set of most relevant documents among which similar/related terms would be added to an original query. This category includes:

- `relevance-feedback` [33], wherein important terms from the top- $k$  retrieved documents are added to the original query based on metrics like `tf-idf`,
- clustering techniques including `termluster` [3], `doccluster` [19], and `conceptcluster` [27], where a graph clustering method like Louvain [2] are employed on a graph whose nodes are the terms and edges are the terms’ pairwise co-occurrence counts so that each cluster would comprise frequently co-occurring terms. Subsequently, to refine the original query, the related terms are chosen from the clusters to which the initial query terms belong.

- `bertqe` [43], which employs `bert`'s contextualized word embeddings of terms in the top- $k$  retrieved documents.

(2) To evaluate whether the expanded gold-standard datasets are indeed effective in improving the performance of supervised models for predicting refined queries, we further establish a benchmark on the generated gold-standard dataset for fine-tuning a pretrained large language model. We opt for text-to-text-transfer-transformer (`t5`) [31], a unified framework to transfer learning for a wide variety of nlp tasks using the same loss function and encoder-decoder architecture by the Transformer [41]. `T5` treats every nlp problem as a text-to-text task, meaning that both the inputs and outputs are always text strings, regardless of the task at hand. This approach allows for a consistent framework that can handle a diverse range of tasks such as translation, summarization, and even more complex tasks like question answering. The model's architecture, based on the Transformer, effectively captures relationships in sequential data, making it highly efficient for natural language understanding and generation tasks. By unifying different NLP tasks under a single model, `t5` simplifies the process of transfer learning and allows for more efficient fine-tuning across tasks, leading to better generalization and performance. It has been pretrained on `c4` large collection of webpages, and, when fine-tuned on benchmark datasets, achieved state-of-the-art performance in text summarization, question answering, and text classification. We fine-tune the base model with 220M parameters for 4,000 epochs on google cloud using `tpus` and use beam search decoding with  $\text{top-}k = 10$  random sampling during inference. We use 70% of  $(q \rightarrow q^*)$  pairs for fine-tuning and evaluate the model's predictions of refined query for the remaining 30% pairs. To provide a minimum base for comparison, we also use pretrained `t5` to generate query refinement without fine-tuning, oblivious to the existing gold-standard datasets.

### 2.3.2 Results

Foremost, due to space constraints, we present only the most significant results in this paper. We refer readers to the codebase for detailed and comprehensive results.

In response to **RQ1**, i.e., whether query backtranslation is effective in scaling up generating gold-standard datasets via producing more refined queries for an original query, from Table 2.3.3, we can observe that query backtranslation can effectively generate more refined queries across *all* query sets, information retrieval methods and evaluation metrics. Specifically, backtranslation showed the best performance on **dbpedia** queries, matching almost half of the original queries with refined versions along with substantial increases in evaluation metrics. This is followed by the **robust04** and **antique** queries, and the poorest performance is associated with **clueweb09b**, which will be discussed in **RQ4** for possible reasons. The latter shows that even in the worst case, there are several refined queries per original query by query backtranslations, which can be used to augment training sets for supervised query refiners. Moreover, from Table 2.3.2, we see that expanded versions of gold-standard datasets using query backtranslation (+bt) consistently boost **t5** performance compared to when it has been trained on datasets generated by only unsupervised baselines, without query backtranslation (-bt). Pretrained **t5** shows the worst performance, which is expected for the model that has not seen any training pairs.

Table 2.3.2: Results of **t5** [31] on gold-standard datasets.

	bm25.map			bm25.ndcg			bm25.mrr		
	t5	t5-fine-tuned		t5	t5-fine-tuned		t5	t5-fine-tuned	
		-bt	+bt		-bt	+bt		-bt	+bt
dbpedia.bm25.map.tsv	0.155	0.325	<b>0.336</b>	0.279	0.496	<b>0.505</b>	0.404	0.768	<b>0.791</b>
robust04.bm25.map.tsv	0.167	0.277	<b>0.286</b>	0.323	0.464	<b>0.475</b>	0.605	0.824	<b>0.841</b>
antique.bm25.map.tsv	0.227	0.488	<b>0.494</b>	0.342	0.591	<b>0.597</b>	0.634	0.972	<b>0.979</b>
gov2.bm25.map.tsv	0.134	0.225	<b>0.228</b>	0.276	0.390	<b>0.393</b>	0.677	0.848	<b>0.869</b>

Table 2.3.3: Efficacy of backtranslated queries in query refinement.  $\#q$  shows the number of original queries that need refinement, while  $\#q^*$  and  $\%$  represent the *best* refined queries’ count and percentage, respectively, and  $\Delta$  denotes the average metric improvements.

		bm25				qld			
		$\#q$	$\#q^*$	$\%$	$\Delta$	$\#q$	$\#q^*$	$\%$	$\Delta$
dbpedia	map	460	192	41.74	+0.11	455	198	43.52	+0.12
	ndcg	461	192	41.65	+0.13	457	195	42.67	+0.13
	mrr	255	140	54.90	+0.44	209	128	61.24	+0.48
robust04	map	249	109	43.78	+0.08	249	105	42.17	+0.07
	ndcg	249	107	42.97	+0.11	249	101	40.56	+0.10
	mrr	112	065	58.04	+0.55	107	068	63.55	+0.49
antique	map	200	060	30.00	+0.07	199	075	37.69	+0.04
	ndcg	200	062	31.00	+0.07	200	081	40.50	+0.06
	mrr	037	019	51.35	+0.60	077	036	46.75	+0.41
gov2	map	149	045	30.20	+0.05	149	041	27.52	+0.06
	ndcg	149	046	30.87	+0.07	149	038	25.50	+0.08
	mrr	057	034	59.65	+0.56	061	026	42.62	+0.58
clueweb09b	map	198	027	13.64	+0.03	198	029	14.65	+0.03
	ndcg	198	027	13.64	+0.05	198	031	15.66	+0.05
	mrr	145	036	24.83	+0.40	163	038	23.31	+0.36

To respond **RQ2**, we compared query backtranslation with global and local unsupervised refiners [38]. In Table 2.3.4, we present the distribution of refined queries over all refiners. As seen, query backtranslation generally outperforms existing unsupervised methods as evidenced by higher counts and percentages of refined queries across different query sets in terms of `map`, and `tagme` and `relevance-feedback` are the runners-up. Similar trends can be observed for `ndcg` and `mrr`, but not presented here for the interest of space. Specifically, as in **RQ1**, query backtranslation shows its best performance in `dbpedia` and `robust04`, finding `clueweb09b`’s queries more challenging for refinement, which is the case for *all* refinement methods and to be

## 2. NO QUERY LEFT BEHIND: QUERY REFINEMENT VIA BACKTRANSLATION

discussed in **RQ4**. Surprisingly, in **antique** query set, **thesaurus** is the best refiner, which can be attributed to the long questions with many terms and the possibility of adding more synonyms overall.

Table 2.3.4: Distribution of refined queries across refinement methods, including query backtranslation, local and global unsupervised refiners in terms of **map**;  $\#q^*$  and  $\%$  show the number of best refined queries and percentage, respectively. **Bold** and underlined numbers are *column-wise* highest and second-highest among refiners, respectively.

		bm25.map										qld.map									
		dbpedia		robust04		antique		gov2		clueweb09b		dbpedia		robust04		antique		gov2		clueweb09b	
		$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$	$\#q^*$	$\%$
Global	backtranslation [ours]	65	<u>13.92</u>	47	<b>18.88</b>	17	8.50	17	<b>11.41</b>	12	6.06	72	<b>15.42</b>	47	<b>18.88</b>	29	<u>14.50</u>	14	9.40	9	4.55
	tagme [10]	70	<b>14.99</b>	19	<u>7.63</u>	19	<u>9.50</u>	13	8.72	22	<b>11.11</b>	62	<u>13.28</u>	20	8.03	17	8.50	12	8.05	19	<b>9.60</b>
	thesaurus [37]	34	7.28	0	0.00	114	<b>57.00</b>	0	0.00	1	0.51	38	8.14	0	0.00	102	<b>51.00</b>	0	0.00	1	0.51
	wiki [36]	26	5.57	16	6.43	1	0.50	7	4.70	9	4.55	18	3.85	18	7.23	1	0.50	11	7.38	15	<u>7.58</u>
	anchor [18]	3	0.64	5	2.01	3	1.50	3	2.01	3	1.52	4	0.86	4	1.61	1	0.50	1	0.67	5	2.53
	conceptnet [8]	10	2.14	12	4.82	2	1.00	6	4.03	5	2.53	13	2.78	12	4.82	2	1.00	4	2.68	4	2.02
	glove [13]	12	2.57	12	4.82	1	0.50	8	5.37	3	1.52	9	1.93	14	5.62	2	1.00	7	4.70	7	3.54
	sense-disambiguation [39]	30	6.42	18	7.23	4	2.00	6	4.03	12	6.06	31	6.64	17	6.83	7	3.50	6	4.03	12	6.06
	word2vec [25]	19	4.07	11	4.42	3	1.50	3	2.01	5	2.53	16	3.43	16	6.43	0	0.00	4	2.68	6	3.03
	wordnet [28]	18	3.85	8	3.21	1	0.50	2	1.34	4	2.02	11	2.36	5	2.01	0	0.00	2	1.34	4	2.02
	stem.krovetz [34]	1	0.21	2	0.80	2	1.00	1	0.67	0	0.00	1	0.21	3	1.20	3	1.50	1	0.67	0	0.00
	stem.lovins [34]	5	1.07	3	1.20	0	0.00	0	0.00	0	0.00	4	0.86	3	1.20	2	1.00	0	0.00	0	0.00
	stem.paicehusk [34]	3	0.64	1	0.40	0	0.00	1	0.67	0	0.00	5	1.07	1	0.40	1	0.50	1	0.67	0	0.00
	stem.porter [34]	2	0.43	2	0.80	11	<u>5.50</u>	0	0.00	0	0.00	1	0.21	1	0.40	1	0.50	0	0.00	0	0.00
	stem.remover [34]	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	stem.trunc4 [34]	1	0.21	1	0.40	1	0.50	0	0.00	0	0.00	2	0.43	2	0.80	0	0.00	0	0.00	1	0.51
	stem.trunc5 [34]	2	0.43	4	1.61	0	0.00	2	1.34	1	0.51	5	1.07	2	0.80	0	0.00	1	0.67	0	0.00
	relevance-feedback [33]	36	7.71	47	<b>18.88</b>	6	3.00	15	<u>10.07</u>	16	<u>8.08</u>	25	5.35	39	<u>15.66</u>	5	2.50	12	8.05	19	<b>9.60</b>
	termcluster [3]	0	0.00	0	0.00	0	0.00	17	<b>11.41</b>	3	1.52	0	0.00	0	0.00	0	0.00	16	<u>10.74</u>	6	3.03
	rm3 [4]	13	2.78	1	0.40	7	3.50	13	8.72	2	1.01	16	3.43	2	0.80	9	4.50	20	<b>13.42</b>	2	1.01
bertqe [43]	5	1.07	3	1.20	0	0.00	1	0.67	2	1.01	1	0.21	1	0.40	2	1.00	0	0.00	4	2.02	
conceptcluster [27]	9	1.93	3	1.20	0	0.00	1	0.67	9	4.55	15	3.21	4	1.61	2	1.00	2	1.34	10	5.05	
doccluster [19]	0	0.00	0	0.00	0	0.00	9	6.04	1	0.51	0	0.00	0	0.00	0	0.00	7	4.70	1	0.51	
hard queries ( $\#q$ )	103	22.06	34	13.65	8	4.00	24	16.11	88	44.44	118	25.27	38	15.26	14	7.00	28	18.79	73	36.87	
total unrefined queries ( $\#q$ )	460	100.00	249	100.00	200	100.00	149	100.00	198	100.00	455	100.00	249	100.00	199	100.00	149	100.00	198	100.00	

For deeper insights, in Figure 2.3.2, we show the distribution of **mrr** improvements between the original query and the refined query by backtranslation and two runner-up methods, i.e., **relevance-feedback**, and **tagme**, across queries. As highlighted, in both the **dbpedia** and **robust04** query sets, backtranslation successfully refined more queries with better **mrr** improvements compared to the other methods. In **clueweb09b**, while most queries are left behind with no refined queries, we can observe that the application of backtranslation has fewer negative impacts.

## 2. NO QUERY LEFT BEHIND: QUERY REFINEMENT VIA BACKTRANSLATION

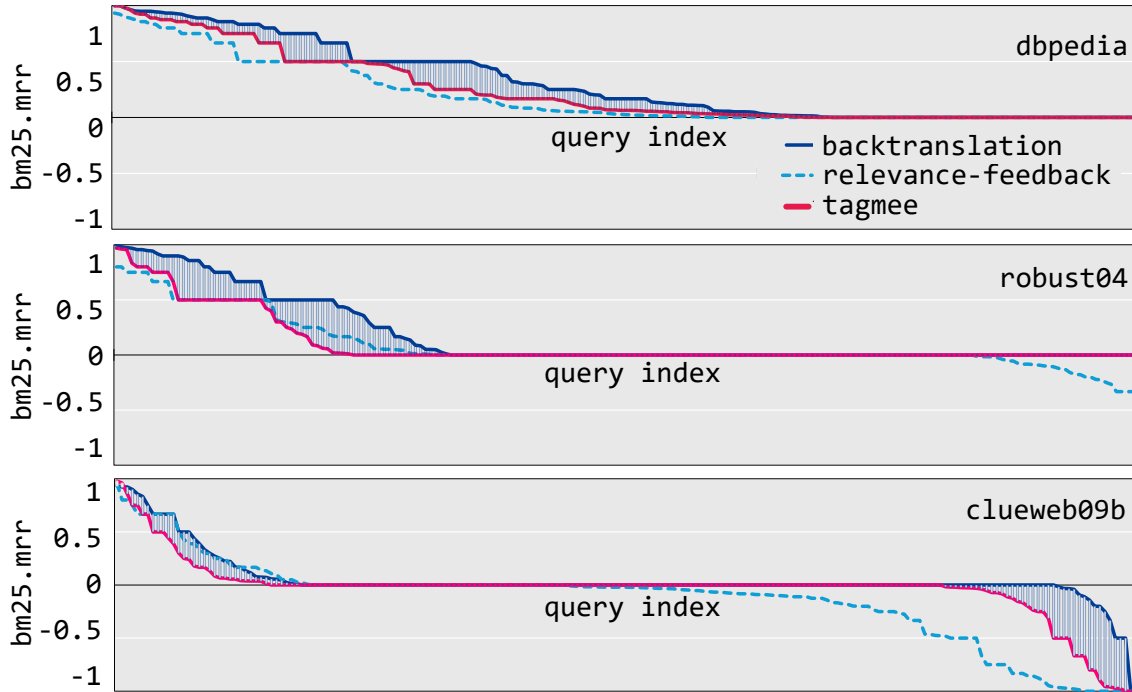


Figure 2.3.2: Distribution of  $\Delta mrr$  across original queries for backtranslation vs. relevance-feedback, and tagmee.

We attribute the superior performance of backtranslation to its ability to introduce diversity and variability into the query space with little to no topic drifts while capturing different aspects of query semantics and nuances in user information needs. From our findings, next to the computational complexity of applying some unsupervised methods such as `bertqe`, we argue that backtranslation represents a valuable lightweight strategy for query refinement.

To answer **RQ3**, i.e., whether backtranslation efficacy is consistent across different languages, looking at Table 2.3.5 and Figure 2.3.3 for `bm25` retriever, we observe that all languages could refine queries, though their efficacy varies. While `arabic` and `swahili` have performed poorly compared to other languages, `chinese`'s performance has been remarkable and consistent across all query sets. It is worth noting that `chinese` belongs to a different language family than `english`, implying that languages from diverse language families are more valuable for reasons like revealing terms that are latent in the source language for being commonly known but should be explicitly mentioned in the target language. Languages of the same family can also be effective

## 2. NO QUERY LEFT BEHIND: QUERY REFINEMENT VIA BACKTRANSLATION

like `russian` and `french`, which are in the same family as `english`, which have demonstrated improvements across nearly all query sets. Since they belong to the same language family, they helped find context-aware synonymous terms and captured the original query’s semantics better. A similar trend is observed in `qld` yet excluded due to space constraints.

Table 2.3.5: Efficacy of query backtranslation across languages; % shows the percentage of queries matched with a refined query, and  $\Delta$  shows the average metric improvements. **Bold** and underlined numbers are *row-wise* highest and second-highest, respectively.

	#q	indo-european						austronesian		dravidian		bantu		sino-tibetan		koreanic		afro-asiatic				
		farsi		french		german		russian		malay		tamil		swahili		chinese		korean		arabic		
		%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	%	$\Delta$	
bm25_map	dbpedia	460	12.17	+0.07	<b>17.17</b>	+0.09	12.61	+0.10	16.30	+0.12	<u>16.96</u>	+0.10	15.00	+0.08	13.48	+0.10	14.57	+0.09	13.26	+0.12	13.26	+0.11
	robust04	249	14.86	+0.05	14.06	+0.06	14.46	+0.06	16.06	+0.06	15.26	+0.05	10.84	+0.09	12.85	+0.06	<u>16.47</u>	+0.05	<b>16.87</b>	+0.06	12.85	+0.09
	antique	200	04.50	+0.07	<u>10.50</u>	+0.06	10.00	+0.05	<b>11.00</b>	+0.05	08.00	+0.03	08.50	+0.04	07.50	+0.06	05.00	+0.04	07.00	+0.05	08.50	+0.04
	gov2	149	09.40	+0.03	09.40	+0.03	9.40	+0.04	<b>10.74</b>	+0.07	08.72	+0.05	08.05	+0.03	06.71	+0.05	<u>10.07</u>	+0.04	<u>10.07</u>	+0.05	06.71	+0.03
	clueweb09b	198	02.53	+0.07	<u>04.04</u>	+0.02	02.53	+0.04	02.53	+0.04	02.53	+0.04	02.02	+0.04	02.02	+0.01	<b>05.56</b>	+0.01	03.03	+0.01	02.53	+0.05
bm25_ndcg	dbpedia	461	13.45	+0.10	<b>17.57</b>	+0.11	13.23	+0.12	16.05	+0.14	<u>17.14</u>	+0.11	14.97	+0.11	13.23	+0.13	14.53	+0.12	11.93	+0.15	14.10	+0.13
	robust04	249	14.46	+0.08	14.46	+0.08	14.06	+0.08	<b>17.27</b>	+0.08	14.46	+0.08	12.45	+0.11	12.85	+0.09	<u>16.47</u>	+0.08	16.06	+0.10	12.05	+0.12
	antique	200	07.50	+0.09	<u>12.00</u>	+0.07	10.00	+0.05	<b>12.50</b>	+0.07	<u>12.00</u>	+0.04	09.50	+0.06	07.50	+0.07	08.00	+0.05	08.00	+0.06	07.50	+0.05
	gov2	149	08.05	+0.04	09.40	+0.04	09.40	+0.07	<u>10.07</u>	+0.07	09.40	+0.07	07.38	+0.03	07.38	+0.04	<b>10.74</b>	+0.04	<b>10.74</b>	+0.06	06.71	+0.04
	clueweb09b	198	02.53	+0.07	<u>03.54</u>	+0.06	01.52	+0.10	2.53	+0.05	01.52	+0.09	02.53	+0.05	0.51	+0.05	<b>05.05</b>	+0.03	03.03	+0.03	02.53	+0.06
bm25_mrr	dbpedia	255	18.43	+0.28	<u>23.53</u>	+0.33	18.82	+0.38	22.35	+0.34	<b>23.92</b>	+0.35	21.18	+0.34	20.00	+0.32	<u>23.53</u>	+0.40	19.22	+0.35	20.00	+0.38
	robust04	112	16.96	+0.44	20.54	+0.32	20.54	+0.42	<b>24.11</b>	+0.44	19.64	+0.47	<b>24.11</b>	+0.43	18.75	+0.37	<u>23.21</u>	+0.40	<u>23.21</u>	+0.32	21.43	+0.38
	antique	037	21.62	+0.40	<u>24.32</u>	+0.50	<u>24.32</u>	+0.35	<b>27.03</b>	+0.52	<b>27.03</b>	+0.53	<u>24.32</u>	+0.35	13.51	+0.40	<b>27.03</b>	+0.54	21.62	+0.43	21.62	+0.48
	gov2	057	14.04	+0.39	21.05	+0.41	15.79	+0.52	14.04	+0.40	14.04	+0.65	17.54	+0.31	12.28	+0.48	<b>28.07</b>	+0.45	<u>22.81</u>	+0.37	15.79	+0.42
	clueweb09b	145	04.14	+0.53	04.83	+0.36	04.14	+0.36	04.83	+0.41	<u>06.21</u>	+0.38	<u>06.21</u>	+0.39	02.76	+0.43	<b>08.28</b>	+0.32	<b>08.28</b>	+0.36	05.52	+0.42

With respect to **RQ4**, from Tables 2.3.3 and 2.3.5, we can observe that query backtranslation can effectively refine queries from a variety of domains overall. However, its efficacy excels in specific domains. As seen, backtranslation demonstrated superior performance in `dbpedia` and `robust04` query sets, and the poorest performance belongs to `clueweb09b`. From Figure 2.3.3, an interesting observation, also relates to **RQ3**, is that while `chinese` and `korean` performed poorly in `antique`, they yield strong results compared to other languages in other query sets. We can see that, in `clueweb09b`, `chinese` reports best results compared to other languages. We attribute the domain-specific performance of languages for query refinement to (1) the queries’ length (number of terms per query) that impacts the quality of backtranslation, and (2) the diversity of topics (genres) in query sets. For the former, Figure 2.3.4

2. NO QUERY LEFT BEHIND: QUERY REFINEMENT VIA BACKTRANSLATION

shows the difference in length of refined vs. original queries across various query sets. As seen, web query sets like `dbpedia` benefit from backtranslated queries, which are long and have more tokens compared to the short and presumably ambiguous original queries; thereby lengthening short queries results in improvement. In contrast, in `antique` where queries are already *long* questions, backtranslated queries that become refined queries yield fewer tokens as they seemingly prune uninformative terms. For the latter, our results show that query refinement via backtranslation for short queries from a general corpus including a wide variety of topics may fall short as in `clueweb09b` compared to long queries from a corpus with a limited span of topics like `dbpedia`.

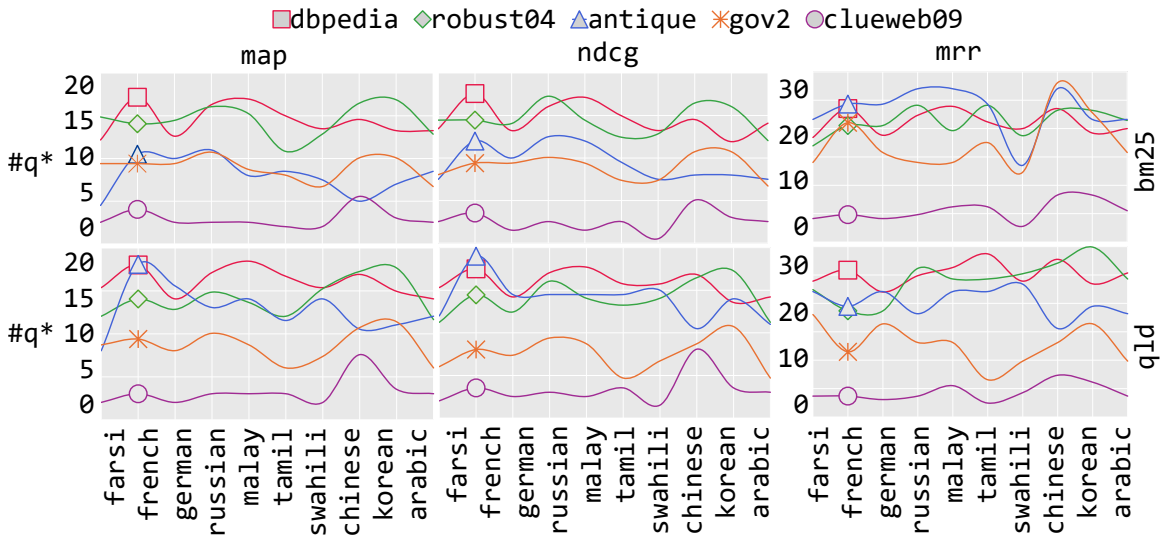


Figure 2.3.3: The language spectrum to illustrate the influence of language across each query set based on the number of *best* refined query obtained by each language.

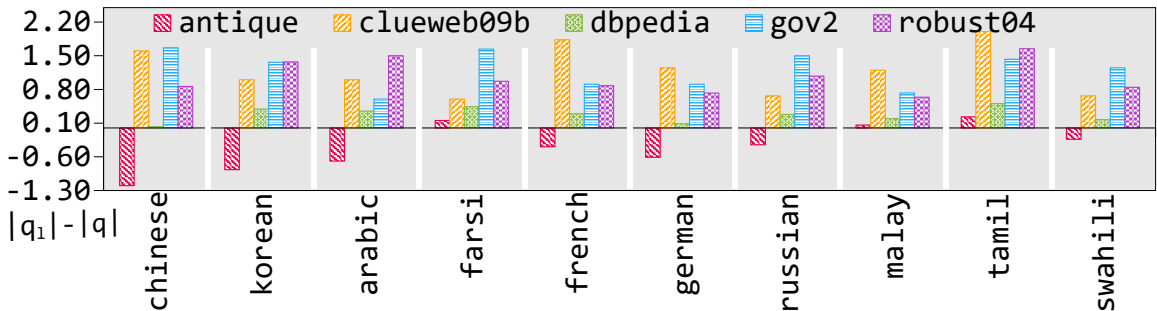


Figure 2.3.4: The length difference between refined query via backtranslation vs. original query.



To answer **RQ5**, i.e., the efficacy of query backtranslation across different translators, Table 2.3.6 shows a comparison between our choice of translator from Meta’s `nllb` [40] and an alternative closed-source translator from Microsoft `bing` [24]. As seen, the application of `nllb` notably yields more refined queries, and `bing` performed poorly. Meanwhile, looking at their translation qualities in Table 3.4.2, we observe that, while both `nllb` and `bing` obtain competitive performance in preserving semantic context in terms of `declutr`, `nllb` yield much diverse with more new terms in backtranslated queries as evidenced by lower values of `rouge-1` compared to `bing`. Table 2.3.6 and Table 3.4.2 together underline that a translator that accurately but with more diverse paraphrases would yield more refined queries.

Table 2.3.6: Meta’s `nllb` vs. Microsoft’s `bing` in query refinement.

		bm25				qld			
		bing		nllb		bing		nllb	
		#q*	%	#q*	%	#q*	%	#q*	%
dbpedia	map	89	19.06	<b>151</b>	<b>32.33</b>	83	17.77	<b>162</b>	<b>34.69</b>
	ndcg	79	16.92	<b>154</b>	<b>32.98</b>	80	17.13	<b>156</b>	<b>33.40</b>
	mrr	41	8.78	<b>129</b>	<b>27.62</b>	35	7.49	<b>117</b>	<b>25.05</b>
robust04	map	49	19.68	<b>87</b>	<b>34.94</b>	46	18.47	<b>89</b>	<b>35.74</b>
	ndcg	43	17.27	<b>87</b>	<b>34.94</b>	45	18.07	<b>87</b>	<b>34.94</b>
	mrr	17	6.83	<b>60</b>	<b>24.10</b>	15	6.02	<b>63</b>	<b>25.30</b>
antique	map	<b>52</b>	<b>26.00</b>	43	21.50	53	26.50	<b>58</b>	<b>29.00</b>
	ndcg	<b>53</b>	<b>26.50</b>	49	24.50	48	24.00	<b>70</b>	<b>35.00</b>
	mrr	7	3.50	<b>19</b>	<b>9.50</b>	11	5.50	<b>34</b>	<b>17.00</b>
gov2	map	26	17.45	<b>37</b>	<b>24.83</b>	30	20.13	<b>31</b>	<b>20.81</b>
	ndcg	22	14.77	<b>40</b>	<b>26.85</b>	22	14.77	<b>33</b>	<b>22.15</b>
	mrr	5	3.36	<b>32</b>	<b>21.48</b>	5	3.36	<b>24</b>	<b>16.11</b>
clueweb09b	map	17	8.59	<b>23</b>	<b>11.62</b>	13	6.57	<b>28</b>	<b>14.14</b>
	ndcg	17	8.59	<b>25</b>	<b>12.63</b>	15	7.58	<b>29</b>	<b>14.65</b>
	mrr	17	8.59	<b>35</b>	<b>17.68</b>	16	8.08	<b>37</b>	<b>18.69</b>

### 2.3.3 Discussion

As the results demonstrate the efficacy of query backtranslation in refining queries across diverse query sets and languages. However, it is essential to note that while backtranslation may face challenges in refining short queries, such as web queries, based on different evaluation metrics, these backtranslated queries can still serve various purposes, such as query suggestions, synonyms, or hints, in other contexts. For instance, as shown in Table 1.1.1, the term `osteoporosis` might pose difficulties for non-native speakers or individuals unfamiliar with technical jargon, but its backtranslated version as `bone disease` can offer them insight into its meaning. Similarly, the term `murals` is backtranslated to `wall paintings`, aiding non-native speakers in comprehending the term.

## 2.4 Concluding Remarks

In this paper, we proposed natural language backtranslation for query refinement to generate gold-standard datasets for supervised query refinement. (1) Our experiments on five query sets, ten languages from varied language families, and two information retrieval methods across three metrics demonstrated the superior performance of query backtranslation against existing unsupervised query refiners. (2) Via fine-tuning `t5` language model on the generated gold-standard datasets with query backtranslations and lack thereof, we showed that the expanded datasets could effectively boost the performance of supervised methods. (3) We further showed that while all languages could match an original query to its refined version, the efficacy rate depends on the choice of language and domain of original query sets. (4) Last, comparing open- and closed-source translators from different platforms, we show that an accurate translator that generates more diverse paraphrases via backtranslation would yield more refined queries. Our future research includes backtranslation *mashup*, i.e., iterative rounds of backtranslation via a mixture of languages.

## References

- [1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. “Dbpedia-a crystallization point for the web of data”. In: *Journal of web semantics* 7.3 (2009), pp. 154–165.
- [2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.
- [3] Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi. “An information-theoretic approach to automatic query expansion”. In: *ACM Transactions on Information Systems (TOIS)* 19.1 (2001), pp. 1–27.
- [4] Marc-Allen Cartright, James Allan, Victor Lavrenko, and Andrew McGregor. “Fast query expansion using approximations of relevance models”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 1573–1576.
- [5] Damir Cavar, Ludovic Mompelat, and Muhammad Abdo. “The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications”. In: *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*. Ed. by Michael Hahn et al. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 46–54. URL: <https://aclanthology.org/2024.sigtyp-1.6>.
- [6] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. “Overview of the TREC 2009 Web Track”. In: *TREC*. 2009.
- [7] Charles LA Clarke, Falk Scholer, and Ian Soboroff. “The TREC 2005 Terabyte Track.” In: *TREC*. 2005.
- [8] *ConceptNet*. <http://conceptnet.io/>.
- [9] Stefano Demichelis and Jörgen W Weibull. “Language, meaning, and games: A model of communication, coordination, and evolution”. In: *American Economic Review* 98.4 (2008), pp. 1292–1311.

- [10] Paolo Ferragina and Ugo Scaiella. “TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)”. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM ’10. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 1625–1628. ISBN: 9781450300995. DOI: 10.1145/1871437.1871689. URL: <https://doi.org/10.1145/1871437.1871689>.
- [11] Angela D Friederici. *Language in our brain: The origins of a uniquely human capacity*. MIT Press, 2017.
- [12] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. “DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, 2021, pp. 879–895. DOI: 10.18653/v1/2021.acl-long.72. URL: <https://doi.org/10.18653/v1/2021.acl-long.72>.
- [13] GloVe. <https://nlp.stanford.edu/projects/glove/>.
- [14] Joan Kelly Hall. *Teaching and researching: Language and culture*. Routledge, 2013.
- [15] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. “ANTIQUÉ: A non-factoid question answering benchmark”. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer. 2020, pp. 166–173.
- [16] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. “DBpedia-entity v2: a test collection for entity search”. In: *Proceedings of the 40th International ACM SI-*

*GIR Conference on Research and Development in Information Retrieval*. 2017, pp. 1265–1268.

- [17] Omar Khattab and Matei Zaharia. “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT”. In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. Ed. by Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu. ACM, 2020, pp. 39–48. DOI: 10.1145/3397271.3401075. URL: <https://doi.org/10.1145/3397271.3401075>.
- [18] Reiner Kraft and Jason Zien. “Mining anchor text for query refinement”. In: *Proceedings of the 13th international conference on World Wide Web*. 2004, pp. 666–674.
- [19] Kyung Soon Lee, W Bruce Croft, and James Allan. “A cluster-based resampling method for pseudo-relevance feedback”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008, pp. 235–242.
- [20] Yuezhong Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia P. Sycara. “Joint Embedding of Hierarchical Categories and Entities for Concept Categorization and Dataless Classification”. In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. Ed. by Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad. ACL, 2016, pp. 2678–2688. URL: <https://www.aclweb.org/anthology/C16-1252/>.
- [21] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: <https://www.aclweb.org/anthology/W04-1013>.
- [22] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. “Pyserini: A Python Toolkit for Reproducible Informa-

- tion Retrieval Research with Sparse and Dense Representations”. In: *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2021, pp. 2356–2362.
- [23] Microsoft. *Azure AI Custom Translator Neural Dictionary Delivering Higher Terminology Translation Quality*. Microsoft. 2023. URL: <https://www.microsoft.com/en-us/translator/blog/2023/12/06/azure-ai-custom-translator-neural-dictionary-delivering-higher-terminology-translation-quality/>.
- [24] Microsoft. *Microsoft Translator GitHub Repository*. <https://github.com/MicrosoftTranslator>. Accessed: [Insert Date].
- [25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).
- [26] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT\* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596. URL: <https://doi.org/10.1145/3287560.3287596>.
- [27] Apostol Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. “Semantic concept-based query expansion and re-ranking for multimedia retrieval”. In: *Proceedings of the 15th ACM international conference on Multimedia*. 2007, pp. 991–1000.
- [28] Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. “Improving query expansion using WordNet”. In: *Journal of the Association for Information Science and Technology* 65.12 (2014), pp. 2469–2478.
- [29] Joao Palotti, Harris Scells, and Guido Zuccon. “TrecTools: an open-source Python library for Information Retrieval practitioners involved in TREC-like campaigns”. In: SIGIR’19. Paris, France: ACM, 2019.

- [30] Jay M Ponte and W Bruce Croft. “A language modeling approach to information retrieval”. In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 202–208.
- [31] Colin Raffel and et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.
- [32] Stephen E. Robertson and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Found. Trends Inf. Retr.* 3.4 (2009), pp. 333–389. DOI: 10.1561/15000000019. URL: <https://doi.org/10.1561/15000000019>.
- [33] Gerard Salton. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971.
- [34] Alexandra Schofield and David Mimno. “Comparing apples to apple: The effects of stemmers on topic models”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 287–300.
- [35] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=B1ckMDqlg>.
- [36] Bashar Al-Shboul and Sung-Hyon Myaeng. “Wikipedia-based query phrase expansion in patent class search”. In: *Information retrieval* 17 (2014), pp. 430–451.
- [37] Ali Asghar Shiri. “End-user interaction with thesaurus-enhanced search interfaces, an evaluation of search term selection for query expansion”. In: (2003).
- [38] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. “Reque: a configurable workflow and dataset collection for query refinement”. In: 2020, pp. 3165–3172.

- [39] Liling Tan. *Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]*. 2014.
- [40] NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. “No language left behind: Scaling human-centered machine translation”. In: *línea*. Disponible en: <https://github.com/facebookresearch/fairseq/tree/nllb> (2022).
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [42] Ellen Voorhees. *Overview of the TREC 2004 Robust Retrieval Track*. en. 2005-08-01 2005. DOI: <https://doi.org/10.6028/NIST.SP.500-261>.
- [43] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. “BERT-QE: Contextualized Query Expansion for Document Re-ranking”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4718–4728. URL: <https://www.aclweb.org/anthology/2020.findings-emnlp.424>.



---

# Chapter 3

## *Enhancing RAG's Retrieval via*

### *Query Backtranslations*

DELARAM RAJAEI, ZAHRA TAHERI, HOSSEIN FANI

Proceedings of the 25th International Web Information Systems Engineering conference (WISE '24), December 2–5, 2024, Qatar University, Doha, Qatar (Submitted)

---

### **3.1 Introduction**

Retrieval-augmented generation (rag) has revolutionized natural language processing by integrating external information retrieval mechanisms to enhance the capabilities of large language models (llms) [21]. By allowing these models to access and incorporate relevant data from external sources in real-time, rag addresses the limitations of static knowledge bases inherent in traditional models. This integration not only improves the accuracy and relevance of generated content but also ensures that responses are informed by the most current information available. As a result, rag significantly reduces the likelihood of producing outdated or inaccurate responses, making it a powerful tool in applications that demand up-to-date knowledge and precise answers. Rag enhances the performance of llms by using two phases: (1) retrieval phase, where it retrieves relevant documents and information related to the original query, and (2) generation phase, where the retrieved documents, combined with the original query, are provided as input to the language model to generate a response. Rag systems have found applications in diverse fields, including product information and customer queries [30], social media by recommending popular hashtags [8], data-to-text generation including drone handover messages [9], and enhancing document

retrieval and robustness [45].

Traditionally, rag systems rely on the input query to retrieve relevant documents from commercial search engines or neural rankers trained on external knowledge source [18]. While effective, this approach may yield limited contextual understanding and suboptimal retrieval outcomes, particularly in addressing complex or diverse informational needs [31]. Rag-fusion advances traditional rag methods by generating variations of the original [30] and considering retrieved lists of documents per query variation, and merging them into a single unified list. This approach enriches the retrieval context, allowing for more comprehensive responses by considering diverse perspectives of user queries [40, 11].

State-of-the-art query expanders are largely based on fine-tuning transformer-based language models [1, 25]. Arabzadeh et al. [1] and others [25] proposed fine-tuning transformer-based language models to generate expanded queries. Zheng et al. [47] leveraged bert to address the introduction of non-relevant information in query expansion. However, fine-tuning such models demands computational resources and time, and has environmental impacts [36]. Moreover, their effectiveness is questionable since evaluation data may have been encountered during pretraining, risking data leakage and overestimating their capabilities [15].

In this research, we propose query backtranslations, that is, translating an original query into other languages and back to the original language, to generate diverse yet contextually relevant query variations [33]. Our proposed query backtranslation is a novel yet simple unsupervised approach, which maintains relevance and controls topic drift. Figure 3.1.1 illustrates the generation of backtranslated versions followed by the fusion of retrieved document sets into a unified set for the generation phase.

## 3.2 Problem Definition

Our goal is to explore the synergistic impact of query backtranslation on generating expanded queries and to utilize reciprocal rank fusion to assess its effect on re-ranking and fusing search results. Herein, we provide a formal statement of the problem in

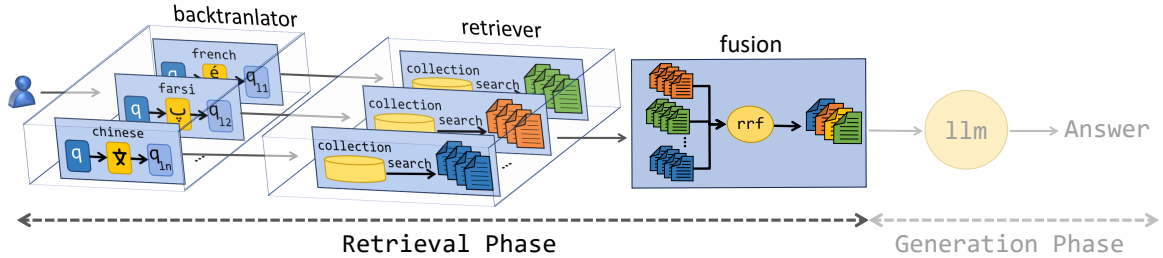


Figure 3.1.1: Generating backtranslated versions of an original query and fusing retrieved document sets for rag-based query refinement.

two steps, after which we propose our methodology.

Given an original query  $q$ , its *retrieved* ranked list of relevant documents  $\mathcal{D}_q$  by a retriever  $r$ , and its *true* list of relevant documents (relevant judgment)  $\mathcal{J}_q$ , our task is to generate  $n$  different versions of  $q$ , denoted by  $\mathcal{Q} = \{q_i\}_{i=1}^n$ , each with its own retrieved ranked list of relevant documents  $\mathcal{D}_{q_i}$  by a retriever  $r$ , such that the reciprocal rank fusion (rrf) [7] of  $n$  ranked lists of  $\mathcal{D}_{q_i}$ ;  $1 \leq i \leq n$ , denoted by  $\mathcal{D}_q^*$  has a better relevance based on  $q$ ’s relevance judgement in terms of an evaluation metric  $m$ .

### 3.3 Proposed Approach

#### 3.3.1 Query Expansion via Natural Language Backtranslation

To generate the variations of an original query, we propose natural language backtranslation. Let  $\mathcal{L}$  be a set of languages. Given a query  $q$ , we backtranslate the query, resulting in a set of modified queries  $q_{\mathcal{L}} = \{q_l : \forall l \in \mathcal{L}\}$ . In our study, without loss of generality to any machine translation models, we leverage meta’s ‘no language left behind’ (nllb)<sup>1</sup>, an open-source neural machine translator capable of providing high-quality translations directly between 200 languages [43]. We opt for nllb for its particular focus on realizing a *universal* translator while prioritizing low-resource natural languages, as opposed to a small dominant subset of natural languages; it

<sup>1</sup>[github.com/facebookresearch/fairseq/tree/nllb](https://github.com/facebookresearch/fairseq/tree/nllb)

enables query backtranslation augmentation via a vast variety of natural languages with distinct properties. Further, `nllb` is open-sourced to foster transparency and can be smoothly integrated into any pipeline with few lines of code.

### 3.3.2 RAG-Based Retrieval

Given the retrieved documents  $\mathcal{D}$  for each backtranslated query  $q_l$  using the information retrieval method  $r$ , we apply reciprocal rank fusion (rrf) [7] to merge outcomes from diverse query versions by considering the ranks of the documents. `rrf` is applied to a given set of retrieved documents  $\mathcal{D}_q^*$  with the constant  $k$ . The constant  $k$  mitigates the impact of high rankings by outlier systems. Formally:

$$\text{rrf}(d \in \mathcal{D}^*) = \sum_{l \in \mathcal{L}} \frac{1}{k + \text{rank}(d)} \quad (1)$$

where  $\text{rank}(d)$  represents the rank of document  $d$  in the list retrieved documents  $\mathcal{D}_{q_l}$  for the query backtranslated version  $q_l$ . The constant  $k$  plays a pivotal role in mitigating the impact of excessively high rankings from outlier systems. Acting as a parameter,  $k$  governs the extent to which these outliers influence the ranking process. Essentially, a higher  $k$  value diminishes the influence of higher rankings, thereby ensuring that the final rankings are less skewed by outliers and better represent the overall quality of the documents retrieved by the refiner. Afterward, all the outcomes are merged and assessed using the metric  $m$ . This metric evaluates the quality of the merged document list for the query  $q$ , denoted as  $m(\mathcal{D}_q^*; \mathcal{J}_q)$ . We select reciprocal rank fusion because while highly ranked documents hold greater significance, the importance of lower-ranked ones should not be disregarded.

## 3.4 Experiment

In this section, we explore the following research questions:

**RQ1:** How does fusion perform across different query reformulation methods? We generate variations of original queries by various unsupervised methods, classified

into four groups. The retrieved document sets for each query version undergo fusion using `rrf` and are evaluated with multiple metrics. This comparison aims to enhance retrieval for improved subsequent generation performance compared to non-fused results.

**RQ2:** Is the effectiveness of `rrf`-fusion consistent across diverse datasets? We compare the fused results from the previous experiment with documents retrieved by the original query across different datasets. These document sets are evaluated using three metrics to gauge their effectiveness.

**RQ3:** What is the impact of the parameter  $k$  on fusion? To examine this parameter, we employ backtranslation in two categories and compute `rrf` with varying values of  $k$ . This analysis investigates how different values of  $k$  influence the fusion results.

### 3.4.1 Dataset

We used well-known query sets in `english` from different domains, namely `dbpedia` [14] collection of wikipedia articles, `robust04` [44] collection of news articles and US government publications, `antique`'s test collection [13] including open-domain non-factoid questions from `Yahoo! Answers`, `gov2` [5] webpages of `.gov` web domain, and `clueweb09b` [4] collection of webpages. In all query sets, we filter out queries with *no* relevance judgment. Also, given an information retrieval method and an evaluation metric, we skip those original queries that result in the best metric value of 1.00, for no refinement is needed. Table 3.4.1 summarizes the statistics of the query sets. As seen in the `robust04`, `gov2`, and `clueweb09b` query sets, the average query lengths are 2.76, 3.13, and 2.45, respectively, indicating relatively short queries. Conversely, the `antique` query set exhibits longer queries, with an average length of 9.34 terms, suggesting more detailed or complex information needs. The `dbpedia` query set falls within an intermediate range, with average query lengths of 5.37 terms.

Table 3.4.1: Query set statistics include query length ( $|q|$ ), the full set of relevant documents ( $\mathcal{J}$ ), and queries.

		avg $m_r(q, \mathcal{J}_q)$							
						bm25		qld	
query set	domain	$\#q$	$\#\text{documents}$	avg $ q $	$ \mathcal{J} $	map	mrr	map	mrr
dbpedia [14]	wikipedia	467	4,635,922	5.37	49,280	0.232	0.565	0.292	0.663
robust04 [44]	news	250	528,155	2.76	311,410	0.199	0.667	0.201	0.681
antique [13]	non-factoid questions	200	403,666	9.34	6,589	0.353	0.881	0.252	0.729
gov2 [5]	*.gov web	150	1,247,753	3.13	135,352	0.157	0.718	0.165	0.706
clueweb09b [4]	web	200	50,000,000	2.45	84,366	0.078	0.383	0.073	0.304

### 3.4.2 Baseline

We compared query backtranslation with 22 existing unsupervised expanders [41] on five datasets. We utilized different unsupervised methods categorized into two groups: local and global. Global methods consider an original query only and include:

- **tagme** [10], which replace the original query’s terms with the title of their **wikipedia** articles,
- **stemmers**, which utilize various lexical, syntactic, and semantic aspects of query terms and their relationships to reduce the terms to their roots, including **krovetz**, **lovins**, **paiceHusk**, **porter**, **sremoval**, **trunc4**, and **trunc5** [35],
- **semantic refiners**, which use an external linguistic knowledge-base including **thesaurus** [39], **wordnet** [27], and **conceptnet** [6], to extract related terms to the original query’s terms,
- **sense-disambiguation** [42], which resolves the ambiguity of polysemous terms in the original query based on the surrounding terms and then adds the synonyms of the query terms as the related terms,
- **embedding-based methods**, which use pre-trained term embeddings from **glove** and **word2vec** [23] to find the most similar terms to the query terms,

- **anchor** [16], which is similar to embedding methods where the embeddings trained on wikipedia articles' *anchors*, presuming an anchor is a concise summary of the content in the linked page,
- **wiki** [38], which uses the embeddings trained on wikipedia's hierarchical categories [19] to add the most similar concepts to each query term.
- **backtranslation**, which a query is translated from its original language (e.g., **english**) to a set of target languages (e.g., **farsi**, **chinese**, ...) from different language families and cultures, including low-resource languages, and then translate it back to the original language.

Local refiners, however, consider terms from top- $k$  retrieved documents via a prior retrieval using an information retrieval method, e.g., **bm25** or **qld**, to find an initial set of most relevant documents among which similar/related terms would be added to an original query. This category includes:

- **relevance-feedback** [34], wherein important terms from the top- $k$  retrieved documents are added to the original query based on metrics like **tf-idf**,
- clustering techniques including **termluster** [3], **doccluster** [17], and **conceptcluster** [26], where a graph clustering method like Louvain [2] are employed on a graph whose nodes are the terms and edges are the terms' pairwise co-occurrence counts so that each cluster would comprise frequently co-occurring terms. Subsequently, to refine the original query, the related terms are chosen from the clusters to which the initial query terms belong.
- **bertqe** [46], which employs **bert**'s contextualized word embeddings of terms in the top- $k$  retrieved documents.

Figure 3.4.1 illustrates the impact of various models on generating variations of the original query. The figure demonstrates a pyramid structure where fine-tuned language models (llms) are positioned towards the bottom. These models tend to produce more specific queries tailored to the datasets they have been trained on. In

contrast, base models, which occupy the top of the pyramid, generate more diverse queries since they are not fine-tuned on any particular datasets. Positioned in the middle are the unsupervised refiners and translator models. Translators, in particular, maintain a balance between diversity and specificity, producing a wide range of queries that remain closely related to the original query without experiencing semantic drift.

### 3.4.3 Setup

Our pipeline involves two stages: 1) generating variations of the original query using query backtranslation and 2) fusion with `rrf` function. Here, we provide the implementation details and the setup of our approach in each of these phases.

#### 3.4.3.1 Query Backtranslation.

We leverage Meta’s ‘*no language left behind*’ (`n11b`) [43]<sup>2</sup>, for being open-source, capable of providing two-way translations in 200 languages with a focus on low-resource languages, and easily integrated into any pipeline with few lines of code. Meta’s `n11b` is available with model card [24] and is developed based on a conditional mixture of several transformers [37] that is trained on data tailored for low-resource languages. On the other extreme, we alternatively chose the `bing` translator<sup>3</sup>, a cloud-based *closed*-source machine translation service offered by Microsoft [22] which supports around 128 languages, yet has *no* publicly available model card and/or documentation, to the best of our search. We deliberately aim to compare the efficacy of our method via two extremes of a well-documented translator against a relatively opaque/obscure translator. We translate queries from `english` into 10 languages from 7 language families, including `malay`, `swahili`, and `tamil` as low-resource languages.

Table 3.4.2 shows the average pairwise similarities between a query and its backtranslated versions using `rouge-1` and `declutr` by Giorgi et al. [12]. Backtranslation from `english` to itself has been performed for unit test purposes where all the re-

<sup>2</sup> <https://github.com/facebookresearch/fairseq/tree/n11b>

<sup>3</sup><https://www.bing.com/Translator>



### 3. ENHANCING RAG’S RETRIEVAL VIA QUERY BACKTRANSLATIONS

sults for `declutr` and `rouge-1` are expected to be the highest possible 1.0 with a negligible change in query length. As seen, all languages could expand the original queries of query sets with new terms in the backtranslated versions with an exception in `antique` set where queries are long questions and backtranslation versions are of the same or contracted lengths, while the semantics remained almost surely intact in terms of `rouge-1` and `declutr` scores. In terms of translation quality, while `rouge-1` considers the overlap of n-grams between a pair of an original and backtranslated query, and hence, falls short of capturing topic drifts, if any, `declutr` relies on the cosine similarity between a pair of query embeddings in a *latent space* and is more effective in measuring semantic similarities. Comparing `nllb` and `bing`, while both translators obtain similar performance in terms of the `declutr`, `bing` has higher values of `rouge-1` indicating *fewer* new terms and *less* diverse paraphrases in backtranslated queries, which yield its poorer performance for query refinement task.

Table 3.4.2: Languages and their families, alongside the translation quality comparison between `nllb` and `bing`. Backtranslation into English is tested to ensure optimal translation quality in the pipeline.

family	language	dbpedia				robust04				antique				gov2				clueweb09b			
		declutr [12]		rouge-1		declutr [12]		rouge-1		declutr [12]		rouge-1		declutr [12]		rouge-1		declutr [12]		rouge-1	
		nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing	nllb	bing
	english	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
indo-european	farsi	0.83	0.85	0.62	0.75	0.81	0.85	0.52	0.72	0.84	0.86	0.63	0.76	0.79	0.86	0.47	0.70	0.74	0.80	0.54	0.73
	french	0.87	0.86	0.70	0.81	0.85	0.86	0.56	0.75	0.89	0.89	0.72	0.81	0.82	0.87	0.52	0.75	0.81	0.83	0.60	0.84
	german	0.85	0.87	0.72	0.83	0.81	0.86	0.54	0.74	0.89	0.89	0.73	0.82	0.79	0.87	0.53	0.73	0.75	0.83	0.59	0.83
	russian	0.86	0.86	0.69	0.79	0.84	0.85	0.56	0.70	0.88	0.86	0.69	0.78	0.81	0.86	0.49	0.68	0.77	0.82	0.54	0.79
austronesian	malay	0.88	0.88	0.69	0.77	0.85	0.88	0.57	0.70	0.88	0.90	0.70	0.81	0.85	0.90	0.53	0.70	0.82	0.84	0.63	0.80
dravidian	tamil	0.84	0.86	0.62	0.81	0.81	0.87	0.50	0.75	0.86	0.87	0.64	0.76	0.82	0.88	0.49	0.79	0.77	0.82	0.56	0.85
bantu	swahili	0.87	0.87	0.69	0.77	0.82	0.86	0.49	0.67	0.88	0.87	0.68	0.76	0.79	0.90	0.44	0.76	0.81	0.84	0.59	0.80
sino-tibetan	chinese	0.80	0.86	0.51	0.71	0.78	0.87	0.45	0.69	0.84	0.86	0.59	0.73	0.77	0.87	0.43	0.64	0.72	0.82	0.42	0.70
koreanic	korean	0.82	0.85	0.58	0.73	0.80	0.84	0.47	0.70	0.83	0.87	0.59	0.75	0.78	0.86	0.43	0.68	0.74	0.81	0.53	0.74
afro-asiatic	arabic	0.83	0.87	0.65	0.77	0.78	0.86	0.53	0.74	0.86	0.87	0.68	0.79	0.77	0.87	0.46	0.69	0.72	0.83	0.51	0.82

#### 3.4.3.2 RAG-Based Retrieval.

We integrated the fusion and re-ranking process, `rrf[7]`<sup>4</sup>, which assigns a reciprocal rank to documents gathered from searches on generated queries from various refiners. We selected this function because it is simpler and more efficient than other

<sup>4</sup><https://github.com/Raudaschl/rag-fusion>

fusion metrics, as it merges ranks without depending on arbitrary scores from specific ranking methods. It functions without requiring a special voting algorithm or global information, allowing ranks to be calculated and combined one system at a time, thus eliminating the need to store all rankings in memory. It utilizes the diversity within individual rankings more effectively, allowing a document, ranked highly by a few systems, to significantly improve its overall rank. Moreover, it prevents a simple majority of weak preferences from overshadowing stronger ones, unlike other fusion metrics [7]. For a more accurate comparison, we calculate the `rrf` metric for groups of documents based on the refiner that generated the query that retrieved the document.

Our approach starts by grouping retrieved documents by `docid` and `qid`. We then iterate through these groups, calculating a relevance score for each document based on its rank within the group. This score incorporates a positive constant  $k$  for normalization or to regulate the impact of rank on the score. We chose to set  $k$  to 60 based on our findings indicating that optimal performance is achieved with a small value, though the specific choice of  $k$  is not crucial.

#### 3.4.4 Search and Evaluation

We have applied two information retrieval methods, namely `bm25` [32] and `qld` [29], using `pyserini` [20] to retrieve relevant content for the original queries as well as the backtranslated versions and evaluate the retrieval performances based on two metrics, i.e., `map`, `mrr`, and `ndcg`, using `trec_eval` [28]. In total, we create a system to retrieve the most relevant documents for the user. A similar trend is observed for `qld`. However, due to space constraints, the results for `qld` can be accessed on our github.

#### 3.4.5 Results

In response to **RQ1**, we addressed the research question by generating various query variations using distinct unsupervised methods. We further fused these variations

according to our five distinct categories: `all` (considering all expanders), `global` (only the global expanders), `local` (only the local expanders), `bt` (backtranslations using `nllb` and `bing` as expanders), and `bt_nllb` (considering backtranslations only from `nllb` translator). This structured approach ensured that we could thoroughly evaluate the performance and efficacy of each refinement method. We evaluate the results of `rrf` and non-fused using `map`, `ndcg`, and `mrr`. Each evaluation was compared against the original query evaluation to identify enhancements. In instances where multiple expanders improved the original query, we only considered the best result among all expanders. Table 3.4.3 represents the results for all five datasets for `bm25.map`. Overall, the `rrf`-based methods exhibit strong performance, with the `rrf.all` category often achieves the highest improvement percentages. This suggests, as expected, that incorporating a diverse set of expanders tends to yield substantial performance gains. It also indicates that a comprehensive approach, combining all expanders enhances the retrieval effectiveness. While `rrf.global` and `rrf.local` also show competitive performance, they are generally outperformed by `rrf.all`, highlighting the advantage of using a holistic set of expanders. Dataset-specific observations further emphasize the benefits of the `rrf.all` approach. For instance, in the `gov2` dataset, `rrf.all` achieves the highest improvement in both metrics, while for the `antique` dataset, `rrf.local` achieves a higher percentage increase in the mentioned metrics, underscoring the effectiveness of localized refiners in certain contexts. As previously mentioned, the `antique` dataset comprises open-domain non-factoid questions, characterized by lengthy queries where each question addresses a specific issue. Local methods enhance these queries by considering terms from the top retrieved documents from an initial retrieval, refining them according to their specific topic. Combining these local methods yields better results than the `all` category. Translation-based expanders, represented by `rrf.bt` and `rrf.bt_nllb`, show less improvement compared to the combined expanders approach. Suggesting that while translation-based refiners contribute positively, their impact may be limited when used in isolation. Therefore, integrating them with other refiners can potentially enhance their effectiveness.

### 3. ENHANCING RAG'S RETRIEVAL VIA QUERY BACKTRANSLATIONS

Table 3.4.3: rrf vs. non-fused results.

		bm25.map									
refiner		dbpedia		robust04		antique		gov2		clueweb09	
		#q**	%	#q**	%	#q**	%	#q**	%	#q**	%
rrf	rrf.all	<b>52</b>	11.13	<u>33</u>	13.25	17	8.50	<b>56</b>	37.58	<b>41</b>	20.81
	rrf.global	44	9.42	18	7.23	18	9.00	7	4.70	<u>25</u>	12.69
	rrf.local	37	7.92	12	4.82	<u>38</u>	19.00	<u>18</u>	12.08	8	4.06
	rrf.bt	21	4.50	9	3.61	0	0.00	8	5.37	6	3.05
	rrf.bt_nllb	12	2.57	11	4.42	0	0.00	1	0.67	6	3.05
global	tagmee	<u>49</u>	10.49	9	3.61	11	5.50	5	3.36	10	5.08
	bt_nllb	40	8.57	27	10.84	8	4.00	7	4.70	9	4.57
	wiki	23	4.93	12	4.82	0	0.00	5	3.36	8	4.06
	thesaurus	22	4.71	0	0.00	<b>72</b>	36.00	0	0.00	0	0.00
	bt_bing	19	4.07	11	4.42	5	2.50	4	2.68	4	2.03
	sensedisambiguation	17	3.64	10	4.02	3	1.50	0	0.00	10	5.08
	word2vec	17	3.64	7	2.81	3	1.50	1	0.67	3	1.52
	wordnet	12	2.57	5	2.01	1	0.50	1	0.67	3	1.52
	conceptnet	9	1.93	9	3.61	1	0.50	4	2.68	5	2.54
	glove	8	1.71	7	2.81	0	0.00	6	4.03	3	1.52
	stem.lovins	3	0.64	3	1.20	0	0.00	0	0.00	0	0.00
	anchor	2	0.43	2	0.80	2	1.00	2	1.34	2	1.02
	stem.porter	2	0.43	1	0.40	4	2.00	0	0.00	0	0.00
	stem.trunc5	2	0.43	3	1.20	0	0.00	2	1.34	1	0.51
	stem.paicehusk	2	0.43	1	0.40	0	0.00	1	0.67	0	0.00
	stem.trunc4	1	0.21	1	0.40	0	0.00	0	0.00	0	0.00
	stem.krovetz	0	0.00	0	0.00	1	0.50	1	0.67	0	0.00
local	relevancefeedback	16	3.43	<b>35</b>	14.06	3	1.50	3	2.01	12	6.09
	rm3	11	2.36	1	0.40	6	3.00	7	4.70	2	1.02
	bertqe	4	0.86	2	0.80	0	0.00	1	0.67	2	1.02
	conceptluster	4	0.86	1	0.40	0	0.00	1	0.67	6	3.05
	docluster	0	0.00	0	0.00	0	0.00	2	1.34	1	0.51
	termluster	0	0.00	0	0.00	0	0.00	5	3.36	2	1.02
<b>q</b>		15	3.21	7	2.81	2	1.00	1	0.67	<u>25</u>	12.69
<b>sum</b>		467	100	249	100	200	100.00	149	100	198	100

### 3. ENHANCING RAG'S RETRIEVAL VIA QUERY BACKTRANSLATIONS

To address **RQ2**, we evaluated the fused results from previous experiments and compared them to documents retrieved by the original query across five datasets. Table 3.4.4 shows the results of comparing our categories with the original. The datasets span different domains, including news articles and non-factoid questions. Across all datasets, the rrf-based methods generally outperformed the original query results. The methods showed a clear trend of higher efficacy, particularly noticeable with the `rrf.all` and `rrf.local`. These categories frequently achieved the highest or second-highest scores across various metrics, indicating an improvement in retrieval performance. When analyzing the performance across different query lengths, the rrf-based methods demonstrated more success with longer queries. In datasets with longer average query lengths, such as `antique`, which has an average query length of 9.34 terms, the improvement was particularly significant. The complexity and detail in longer queries benefited more from the diverse retrieval approaches of the rrf methods. In contrast, for datasets with shorter average query lengths, such as `robust04` (average query length of 2.76 terms) and `clueweb09b` (average query length of 2.45 terms), the improvement was present but less pronounced. The shorter queries, which are often more straightforward, did not leverage the full potential of the rrf-based methods as effectively as longer, more complex queries did.

Table 3.4.4: Comparison of the efficacy of rrf-based and original rrf-based results across different datasets.

		dbpedia			robust04			antique			gov2			clueweb09		
		#q*	%	avg	#q*	%	avg	#q*	%	avg	#q*	%	avg	#q*	%	avg
bm25.map	original	23	4.93	0.232	14	5.62	0.199	9	4.50	0.353	1	0.67	0.157	29	14.65	0.078
	rrf.all	<b>96</b>	20.56	0.289	<b>62</b>	24.90	0.223	<u>37</u>	18.50	0.404	<b>71</b>	47.65	0.231	<b>62</b>	31.31	0.088
	rrf.global	<u>88</u>	18.84	0.241	38	15.26	0.211	24	12.00	0.350	14	9.40	0.167	<u>39</u>	19.70	0.057
	rrf.local	87	18.63	0.210	<u>46</u>	18.47	0.183	<b>107</b>	53.50	0.239	<u>36</u>	24.16	0.131	21	10.61	0.051
	rrf.bt	48	10.28	0.258	22	8.84	0.220	1	0.50	0.446	17	11.41	0.214	13	6.57	0.065
	rrf.bt_nllb	28	6.00	0.234	19	7.63	0.197	1	0.50	0.240	4	2.68	0.164	14	7.07	0.067

Comparing the retrieved documents from the these methods to those from the original query, the rrf-based methods consistently retrieved more relevant documents and achieved higher average scores. This improvement suggests that the rrf approach

### 3. ENHANCING RAG'S RETRIEVAL VIA QUERY BACKTRANSLATIONS

provides a more detailed and comprehensive retrieval process, capturing a broader range of relevant information. Among the different categories, `rrf.all` emerged as the most successful. This category, which considers all documents retrieved for all query variations, consistently achieved the highest scores across various metrics. The broad and inclusive nature of this method likely contributed to its success, as it combines the strengths of multiple query expansions and retrieval strategies, leading to a more effective overall retrieval process.

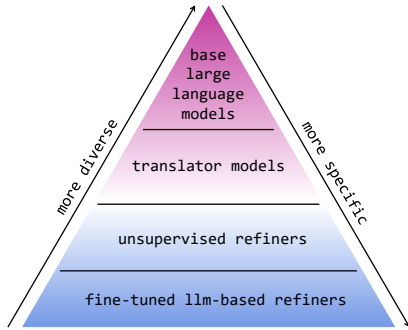


Figure 3.4.1: Effect of different models on refining queries.

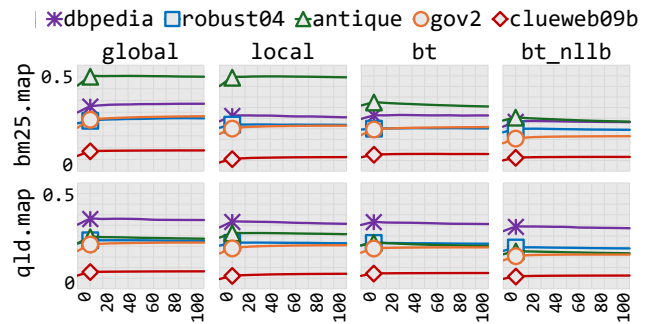


Figure 3.4.2: Effect of constant  $k$  range on fusion outcomes across various categories.

To answer **RQ3** and observe the effect of the constant  $k$  in the `rrf`, we conducted multiple experiments across different values in  $\{0, 10, 20, \dots, 100\}$ . From Figure 3.4.2 as expected from the results of `rrf`, the experiments indicated that  $k$  equal to 60 was near-optimal, though the choice of  $k$  was not critically sensitive. This suggests that while  $k$  is an important parameter, the robustness of `rrf` in providing high-quality rankings remains consistent across a range of  $k$  values, reinforcing its utility in various contexts. Essentially, a higher  $k$  value reduces the influence of higher rankings, thereby ensuring that the final rankings are less skewed by outliers and better represent the overall quality of the documents retrieved by the refiner. Experiential observations suggest that  $k$  performs best when set to a small value, such as 60[7]. Note that this  $k$  value is a constant in the `rrf` algorithm and is entirely distinct from the  $k$  that regulates the number of nearest neighbors.

## 3.5 Concluding Remarks

In this research, we proposed backtranslation as an unsupervised method to enhance the retrieval phase of retrieval-augmented generation (rag) systems. We showed that query backtranslation creates diverse and semantically enriched variations of the original query without semantic drift and, hence, could improve the retrieval phase of rag systems. Our experiment demonstrated that (1) fusion methods generally outperform other query reformulation methods. Specifically, query backtranslation demonstrated substantial performance gains. (2) The efficacy of `rrf` is consistent across diverse datasets, and (3) `rrf` consistently provides high-quality rankings across a range of values for its hyperparameter. Our future research includes studying the effect of these improved retrieved documents on the generation phase. Further, we will explore the effectiveness of additional fusion metrics such as `combmz` [7].

## References

- [1] Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. “Matches Made in Heaven: Toolkit and Large-Scale Datasets for Supervised Query Reformulation”. In: 2021, pp. 4417–4425.
- [2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.
- [3] Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi. “An information-theoretic approach to automatic query expansion”. In: *ACM Transactions on Information Systems (TOIS)* 19.1 (2001), pp. 1–27.
- [4] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. “Overview of the TREC 2009 Web Track”. In: *TREC*. 2009.
- [5] Charles LA Clarke, Falk Scholer, and Ian Soboroff. “The TREC 2005 Terabyte Track.” In: *TREC*. 2005.
- [6] *ConceptNet*. <http://conceptnet.io/>.
- [7] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. “Reciprocal rank fusion outperforms condorcet and individual rank learning methods”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 758–759.
- [8] Run-Ze Fan, Yixing Fan, Jianguai Chen, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. “RIGHT: Retrieval-Augmented Generation for Mainstream Hashtag Recommendation”. In: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I*. Ed. by Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis. Vol. 14608. Lecture Notes in Computer Science. Springer, 2024, pp. 39–55. DOI: 10.1007/978-3-031-56027-9\_3. URL: [https://doi.org/10.1007/978-3-031-56027-9\\_3](https://doi.org/10.1007/978-3-031-56027-9_3).



- [9] Ruitao Feng, Xudong Hong, Mayank Jobanputra, Mattes Warning, and Vera Demberg. “Retrieval-Augmented Modular Prompt Tuning for Low-Resource Data-to-Text Generation”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Ni-anwen Xue. ELRA and ICCL, 2024, pp. 14053–14062. URL: <https://aclanthology.org/2024.lrec-main.1224>.
- [10] Paolo Ferragina and Ugo Scaiella. “TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)”. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10*. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 1625–1628. ISBN: 9781450300995. DOI: 10.1145/1871437.1871689. URL: <https://doi.org/10.1145/1871437.1871689>.
- [11] Haomin Fu, Yeqin Zhang, Haiyang Yu, Jian Sun, Fei Huang, Luo Si, Yongbin Li, and Cam-Tu Nguyen. “Doc2Bot: Accessing Heterogeneous Documents via Conversational Bots”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 2022, pp. 1820–1836. DOI: 10.18653/v1/2022.FINDINGS-EMNLP.131. URL: <https://doi.org/10.18653/v1/2022.findings-emnlp.131>.
- [12] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. “DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, 2021, pp. 879–895. DOI: 10.18653/

v1/2021.acl-long.72. URL: <https://doi.org/10.18653/v1/2021.acl-long.72>.

- [13] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. “ANTIQUE: A non-factoid question answering benchmark”. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer. 2020, pp. 166–173.
- [14] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. “DBpedia-entity v2: a test collection for entity search”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2017, pp. 1265–1268.
- [15] Xing Hu, Ling Liang, Xiaobing Chen, Lei Deng, Yu Ji, Yufei Ding, Zidong Du, Qi Guo, Timothy Sherwood, and Yuan Xie. “A Systematic View of Model Leakage Risks in Deep Neural Network Systems”. In: *IEEE Trans. Computers* 71.12 (2022), pp. 3254–3267. DOI: 10.1109/TC.2022.3148235. URL: <https://doi.org/10.1109/TC.2022.3148235>.
- [16] Reiner Kraft and Jason Zien. “Mining anchor text for query refinement”. In: *Proceedings of the 13th international conference on World Wide Web*. 2004, pp. 666–674.
- [17] Kyung Soon Lee, W Bruce Croft, and James Allan. “A cluster-based resampling method for pseudo-relevance feedback”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008, pp. 235–242.
- [18] Patrick S. H. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*. Ed. by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020. URL:

<https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481Abstract.html>.

- [19] Yue Zhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia P. Sycara. “Joint Embedding of Hierarchical Categories and Entities for Concept Categorization and Dataless Classification”. In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. Ed. by Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad. ACL, 2016, pp. 2678–2688. URL: <https://www.aclweb.org/anthology/C16-1252/>.
- [20] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. “Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations”. In: *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2021, pp. 2356–2362.
- [21] Nelson F. Liu, Tianyi Zhang, and Percy Liang. “Evaluating Verifiability in Generative Search Engines”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics, 2023, pp. 7001–7025. DOI: 10.18653/v1/2023.FINDINGS-EMNLP.467. URL: <https://doi.org/10.18653/v1/2023.findings-emnlp.467>.
- [22] Microsoft. *Azure AI Custom Translator Neural Dictionary Delivering Higher Terminology Translation Quality*. Microsoft. 2023. URL: <https://www.microsoft.com/en-us/translator/blog/2023/12/06/azure-ai-custom-translator-neural-dictionary-delivering-higher-terminology-translation-quality/>.
- [23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).
- [24] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Ge-

- bru. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT\* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596. URL: <https://doi.org/10.1145/3287560.3287596>.
- [25] Yogeswar Lakshmi Narayanan and Hossein Fani. “RePair: An Extensible Toolkit to Generate Large-Scale Datasets via Transformers for Query Refinement”. In: *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management, University of Birmingham and Eastside Rooms, UK, October 21-25, 2023*. ACM, 2023. DOI: 10.1145/3583780.3615129.
- [26] Apostol Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. “Semantic concept-based query expansion and re-ranking for multimedia retrieval”. In: *Proceedings of the 15th ACM international conference on Multimedia*. 2007, pp. 991–1000.
- [27] Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. “Improving query expansion using WordNet”. In: *Journal of the Association for Information Science and Technology* 65.12 (2014), pp. 2469–2478.
- [28] Joao Palotti, Harrisen Scells, and Guido Zuccon. “TrecTools: an open-source Python library for Information Retrieval practitioners involved in TREC-like campaigns”. In: SIGIR’19. Paris, France: ACM, 2019.
- [29] Jay M Ponte and W Bruce Croft. “A language modeling approach to information retrieval”. In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 202–208.
- [30] Zackary Rackauckas. “RAG-Fusion: a New Take on Retrieval-Augmented Generation”. In: *CoRR* (2024).
- [31] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. “The Curious Case of Hallucinations in Neural Machine Translation”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, On-*

- line*, June 6-11, 2021. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Association for Computational Linguistics, 2021, pp. 1172–1183. DOI: 10.18653/V1/2021.NAAACL-MAIN.92. URL: <https://doi.org/10.18653/v1/2021.naacl-main.92>.
- [32] Stephen E. Robertson and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Found. Trends Inf. Retr.* 3.4 (2009), pp. 333–389. DOI: 10.1561/15000000019. URL: <https://doi.org/10.1561/15000000019>.
- [33] Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. “End-to-End Training of Neural Retrievers for Open-Domain Question Answering”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, 2021, pp. 6648–6662. DOI: 10.18653/V1/2021.ACL-LONG.519. URL: <https://doi.org/10.18653/v1/2021.acl-long.519>.
- [34] Gerard Salton. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971.
- [35] Alexandra Schofield and David Mimno. “Comparing apples to apple: The effects of stemmers on topic models”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 287–300.
- [36] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. “Green ai”. In: *Communications of the ACM* 63.12 (2020), pp. 54–63.
- [37] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Con-*

- ference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=B1ckMDqlg>.
- [38] Bashar Al-Shboul and Sung-Hyon Myaeng. “Wikipedia-based query phrase expansion in patent class search”. In: *Information retrieval 17* (2014), pp. 430–451.
- [39] Ali Asghar Shiri. “End-user interaction with thesaurus-enhanced search interfaces, an evaluation of search term selection for query expansion”. In: (2003).
- [40] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. “Retrieval Augmentation Reduces Hallucination in Conversation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, 2021, pp. 3784–3803. DOI: 10.18653/V1/2021.FINDINGS-EMNLP.320. URL: <https://doi.org/10.18653/v1/2021.findings-emnlp.320>.
- [41] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. “Reque: a configurable workflow and dataset collection for query refinement”. In: 2020, pp. 3165–3172.
- [42] Liling Tan. *Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]*. 2014.
- [43] NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. “No language left behind: Scaling human-centered machine translation”. In: *línea*. Disponible en: <https://github.com/facebookresearch/fairseq/tree/nllb> (2022).
- [44] Ellen Voorhees. *Overview of the TREC 2004 Robust Retrieval Track*. en. 2005-08-01 2005. DOI: <https://doi.org/10.6028/NIST.SP.500-261>.

- [45] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. “Corrective Retrieval Augmented Generation”. In: *CoRR* abs/2401.15884 (2024). DOI: 10.48550/ARXIV.2401.15884. arXiv: 2401.15884. URL: <https://doi.org/10.48550/arXiv.2401.15884>.
- [46] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. “BERT-QE: Contextualized Query Expansion for Document Re-ranking”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4718–4728. URL: <https://www.aclweb.org/anthology/2020.findings-emnlp.424>.
- [47] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. “BERT-QE: contextualized query expansion for document re-ranking”. In: *arXiv preprint arXiv:2009.07258* (2020).

---

## Chapter 4

### *Poster Presentations*

---


To showcase the practical implications of our study, we engaged in the *University of Windsor's 9th Annual Computer Science Demo Day*. During this event, we had the opportunity to discuss and exhibit the practical applications of our research project to professionals from various sectors of the technology industry. We opted for a poster presentation format, recognizing its strengths in visual communication and its effectiveness in capturing the interest of attendees. The following sections of this chapter will feature the posters that were displayed.



## 4.1 University of Windsor's 9th Demo Day

Matches  
Made in  
Heaven

Or Somewhere



**Introduction**

Web queries frequently present challenges as they tend to be concise and vague due to users' uncertainty in expressing their information needs. The retrieval of relevant search results could be enhanced by modifying user's initial queries considering user's context.

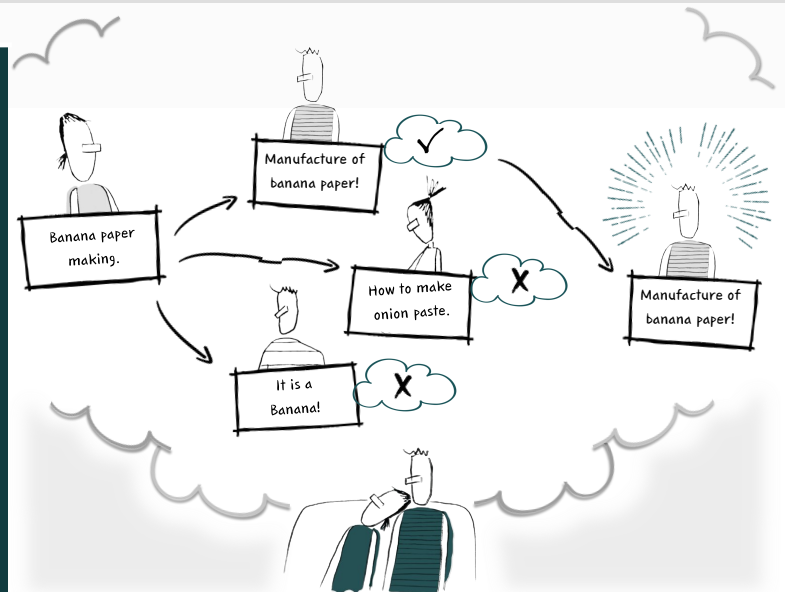
**Methodology**

Our software workflow takes three inputs:

- 1) Dataset of queries with relevance judgments
- 2) Information retrieval method
- 3) Evaluation metric


**Results**

- Generated and shared gold standard datasets
- Publicly available source code
- Availability of gold datasets and benchmark results
- Evaluation of state-of-the-art supervised and unsupervised query refinement methods



**Links**

<https://github.com/fani-lab/RePair>





**Authors**

Delaram Rajaei, Zahra Taherikhonakdar, Michele Catani, Mukesh Reddy Somireddy, Dr. Hossein Fani

**Mentor**

Dr. Hossein Fani

University of Windsor

Figure 4.1.1: The poster we presented at University of Windsor's 9th Annual Computer Science Demo Day

---

# Chapter 5

## *Conclusion*

---

### 5.1 Research Questions

This section presents the details of research questions that we answered through this thesis.

**RQ1**, delves into the implementation of backtranslation using 10 languages across 7 language families, including low-resource languages, as a refinement technique within our pipeline. Our evaluation focused on determining how many backtranslated queries became refined queries and the extent to which they improved each evaluation metric. We further tested the effectiveness of the scale-up for supervised methods by fine-tuning a large language model with the generated datasets, both with and without backtranslations. The results, as presented in Table 2.3.3, indicate that query backtranslation effectively generates more refined queries across all query sets, information retrieval methods, and evaluation metrics. Specifically, the best performance was observed with `dbpedia` queries, where nearly half of the original queries were matched with refined versions, along with substantial improvements in evaluation metrics. Notably, even in the worst-case scenario, several refined queries per original query were generated through backtranslation, which can augment training sets for supervised query refiners. Additionally, as shown in Table 2.3.2, the expanded gold-standard datasets using query backtranslation (+bt) consistently boosted the performance of `t5` compared to datasets generated without query backtranslation (-bt). The pretrained `t5` showed the worst performance, as expected, since it had

not been exposed to any training pairs.

**RQ2**, compared refined queries resulting from backtranslation against 22 unsupervised refiners across various information retrieval methods, evaluation metrics, and query sets from different domains. Our findings, summarized in Table 2.3.4, indicate that query backtranslation generally outperforms existing unsupervised methods, as evidenced by higher counts and percentages of refined queries across different query sets in terms of mean average precision (**map**). Specifically, backtranslation showed its best performance with the **dbpedia** and **robust04** query sets, while **clueweb09b** queries were more challenging for refinement for all methods. As illustrated in Figure 2.3.2, backtranslation achieved better **mrr** improvements compared to **relevance-feedback** and **tagme**, particularly in the **dbpedia** and **robust04** query sets. Although **clueweb09b** queries often remained unrefined, backtranslation had fewer negative impacts. We attribute the superior performance of backtranslation to its ability to introduce diversity and variability into the query space without significant topic drift, effectively capturing different aspects of query semantics and user information needs. We conclude that backtranslation represents a valuable lightweight strategy for query refinement.

**RQ3** conducted a comparative analysis of languages from 7 different language families to determine whether the semantic coherence of backtranslated queries is influenced by the linguistic relationship between the source and target languages. Our hypothesis was that source and target languages from the same family would produce more semantically related queries, while those from different families might generate more diverse outputs. As shown in Table 2.3.5 and Figure 2.3.3 for the **bm25** retriever, we found that all languages could refine queries, but their efficacy varied. Notably, it also suggests that languages from diverse families can be valuable in revealing latent terms in the source language that need explicit mention in the target language. Conversely, languages within the same family, like which share a family with **english**, also demonstrated significant improvements, highlighting their ability to find context-aware synonyms and better capture the original query’s semantics.

**RQ4** generated query backtranslations for 5 query sets with varying query lengths

and topics from different domains, such as news articles versus web content. From Tables 2.3.3 and 2.3.5, we observed that query backtranslation effectively refines queries across various domains, though its efficacy is more pronounced in specific domains. We attribute this domain-specific performance to two main factors: the length of the queries and the diversity of topics. Figure 2.3.4 illustrates that web query sets like `dbpedia`, which benefit from longer backtranslated queries with more tokens, show significant improvements. In contrast, in the `antique` set, where queries are already long, backtranslated queries tend to prune uninformative terms, resulting in fewer tokens. Additionally, query refinement via backtranslation for short queries from a general corpus with diverse topics is less effective compared to long queries from a corpus with a limited span of topics.

**RQ5** conducted experiments using two neural machine translators built on different technologies and platforms: Meta’s `nllb` and Microsoft’s `bing`. As shown in Table 2.3.6, `nllb` produced notably more refined queries compared to `bing`, which performed poorly. Examining their translation qualities in Table 3.4.2, we observed that while both achieved competitive performance in preserving semantic context, `nllb` generated more diverse outputs with a greater number of new terms in backtranslated queries, evidenced by lower `rouge-1` values compared to `bing`. The combined results from Table 2.3.6 and Table 3.4.2 suggest that a translator that provides accurate yet diverse paraphrases yields more effective refined queries. Thus, the choice of neural machine translator significantly impacts the efficacy of query backtranslation.

## 5.2 Concluding Remarks and Future Work

In this thesis, we proposed natural language backtranslation for query refinement to generate gold-standard datasets for supervised query refinement. Our experiments on five query sets, ten languages from varied language families, and two information retrieval methods across three metrics demonstrated the superior performance of query backtranslation against existing unsupervised query refiners. By fine-tuning the `t5` language model on the generated gold-standard datasets with and without query

backtranslations, we showed that the expanded datasets could effectively boost the performance of supervised methods. Additionally, we found that while all languages could match an original query to its refined version, the efficacy rate depends on the choice of language and the domain of the original query sets. Lastly, comparing open- and closed-source translators from different platforms, we demonstrated that an accurate translator that generates more diverse paraphrases via backtranslation would yield more refined queries.

Furthermore, we explored the application of backtranslation as an unsupervised method to enhance the retrieval phase of retrieval-augmented generation (rag) systems. We showed that query backtranslation creates diverse and semantically enriched variations of the original query without semantic drift, thereby improving the retrieval phase of rag systems. Our experiments demonstrated that fusion methods generally outperform other query reformulation methods, with query backtranslation showing substantial performance gains. The efficacy of reciprocal rank fusion (rrf) was consistent across diverse datasets, consistently providing high-quality rankings across a range of values for its hyperparameter. Looking forward, our future endeavors aim to investigate the effect of these improved retrieved documents on the generation phase and backtranslation mashups, involving iterative rounds of backtranslation through a mixture of languages.

# VITA AUCTORIS

NAME: Delaram Rajaei

PLACE OF BIRTH: Tehran, Iran

YEAR OF BIRTH: 1999

EDUCATION: Amirkabir University of Technology  
(Tehran Polytechnic), B.Sc in Computer  
Engineering, Tehran, Iran, 2023

University of Windsor, M.Sc in Computer  
Science, Windsor, Ontario, 2024