No Query Left Behind: Query Refinement via Backtranslation

Delaram Rajaei
دلارام رجایی
School of Computer Science
University of Windsor, ON., Canada
rajaeid@uwindsor.ca

Zahra Taheri
زهرا طاهری
School of Computer Science
University of Windsor, ON., Canada
taherik@uwindsor.ca

Hossein Fani
حسین فانی
School of Computer Science
University of Windsor, ON., Canada
hfani@uwindsor.ca

Abstract

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 → 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'*→(`tamil`) 'அத்திமரங்கள்'→ *'the fig trees'*), (2) augment a query with context-aware synonyms in a target language (e.g., *'italian nobel prize winners'*→(`farsi`) 'برنده های جایزه نوبل ایتالیایی'→ *'italian nobel laureates'*, and (3) help with the semantic disambiguation of polysemous terms and collocations (e.g., *'custer's last stand'* →(`malay`) *'pertahan terakhir custer'*→ *'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`.

CCS Concepts

• **Information systems → Query reformulation**; • **Computing methodologies → Machine translation**.

Keywords

Query Reformulation; Backtranslation; Gold-Standard Generation;

ACM Reference Format:
Delaram Rajaei, Zahra Taheri, and Hossein Fani. 2024. No Query Left Behind: Query Refinement via Backtranslation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627673.3679729

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679729

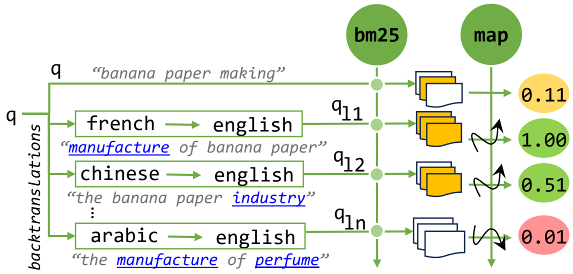

Figure 1: Query backtranslation workflow.

1 Introduction

Retrieving relevant information poses challenges to search engines when user queries are short and unclear, leading to the retrieval of *ir*relevant 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 [4, 60] or seq-to-seq encoder-decoder neural architecture [5, 22], 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 [22]. However, users' intent may undergo gradual or sudden changes in topics within a search session intrinsically by e.g., search engine's *in*correct suggestion of unrelated terms [71], or extrinsically by e.g., online ads, resulting in a loss of sequential semantic context between queries, known as *topic (query) drift* [19, 71]. 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 [10].

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 [8, 81, 96]. Tamannaee et al. [81] 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. [8] and others [62] proposed

Table 1: Queries and the efficacy of their backtranslations.

query id	original query (q)	(language) translation	backtranslation (q_l)	map$_{q_l}$ (Δ_{q_l-q})
dbpedia				
SemSearch_ES-13	*banana paper making*	(korean) 바나나 종이 제조	*manufacture of banana paper*	1.00 (+0.89)
INEX_XER-116	*italian nobel prize winners*	(farsi) برنده های ایتالیایی جایزه نوبل	*italian nobel laureates*	0.57 (+0.34)
INEX_LD-2010057	*einstein relativity theory*	(swahili) *nadharia ya uhusiano wa einstein*	*einstein theory of relation*	0.01 (-0.30)
robust04				
314	*marine vegetation*	(chinese) 海生植物	*the seaweed*	0.19 (+0.19)
426	*law enforcement, dogs*	(swahili) *polisi, mbwa*	*police dogs*	0.33 (+0.29)
338	*risk of aspirin*	(arabic) خطر الأسبرين	*the dangers of aspirin*	0.15 (-0.25)
antique				
421753	*how to get rid of a skunk?*	(swahili) *jinsi ya kuondoa skunk?*	*how to remove skunk*	0.25 (+0.05)
1702151	*how patient a driver are you?*	(french) *Vous êtes un chauffeur patient ?*	*are you a patient driver?*	0.35 (+0.12)
204633	*why do you have memories?*	(korean) 왜 기억이 나나요?	*why do you remember*	0.00 (-0.11)
gov2				
804	*ban on human cloning*	(farsi) ممنوعیت کلون کردن انسان	*pohibition of human cloning*	1.00 (+0.48)
822	*custer's last stand*	(malay) *pertahan terakhir custer*	*custer's last defense*	0.13 (+0.03)
753	*bullying prevention programs*	(french) *programmes de prévention de l'intimidation*	*the prevention of bullying programmes*	0.06 (-0.03)
clueweb09b				
154	*figs*	(tamil) அத்திமரங்கள்	*the fig trees*	1.00 (+0.91)
130	*fact on uranus*	(korean) 천왕성에 대한 사실	*the facts about uranus*	0.16 (+0.01)
51	*horse hooves*	(farsi) کفش اسب	*horse shoes*	0.03 (-0.19)

fine-tuning transformer-based language models to generate (query → refined query) pairs. Fine-tuning a transformer, however, demands significant computational resources and time along with its environmental impact [74]. 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 [37, 90].

In this paper, 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 [84]) 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 brain [29], 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 [95]:

- 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 [18]. 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, 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 [78]. 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 *'custer's last stand'*[1] to malay *'pertahan terakhir custer'*, and backtranslating to english, *'custer s 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 [27, 36, 50, 92] and other natural language processing tasks like text summarization [26] and question-answering [9], and machine translation [31, 47, 75]. Furthermore, the open-source accessibility to multilingual neural machine translators [42, 84, 91], 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 paper, we proposed a reproducible domain-agnostic pipeline to generate refined queries via language backtanslation. From Figure 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 ($q \rightarrow q^\star$) such that $q^\star$ retrieves better search results compared to q under the information retrieval method and the evaluation metric. Our findings show that query backtranslation substantially expands 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 backtranslations has further been evidenced via the performance boost of a fine-tuned large language

[1] https://en.wikipedia.org/wiki/battle_of_the_little_bighorn

model (t5 [68]). 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 contributions lie on:

(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[2], 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 [68], 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.

2 Related Work

The work related to this paper can be broadly categorized into (1) query refinement methods and (2) backtranslation applications.

2.1 Query Refinement

Proposed methods for query refinement, variously referred to by such other names as query rewriting, query reformulation, or query expansion, are either unsupervised, supervised, or semi-supervised. Earlier approaches were mostly unsupervised and involved modifying an original query by expanding and/or replacing query terms based on synonyms from an external source like a thesaurus [40, 44, 78, 79] or based on inter-term correlations or cooccurrences within a training corpus [54, 59]. Unsupervised methods, however, overlook the query's semantic context and may replace polysemous terms with terms that yield topic drift. To control semantic drift, Rocchio [70] and others [39, 70] proposed to modify the query based on terms in the set of clicked documents as relevance feedback from users. In the absence of user feedback, pseudo-relevance-feedback [13, 14, 83, 93] were proposed to modify the query based on the top few documents retrieved by a search engine or an information retrieval method.

Successful as they are for short queries, unsupervised methods fall short for detailed and long queries. To fill the gap, semi-supervised and supervised techniques were proposed to learn users' intents from users' search logs and generate a refined query by considering semantic and contextual aspects of users' search sessions [5, 15, 22, 24, 33, 46, 53, 80, 94, 97, 100]. Sordoni et al. [80] proposed a hierarchical recurrent encoder-decoder architecture to first encode the sequence of terms at the query level using a unidirectional recurrent neural network. Next, a unidirectional recurrent neural network encodes the search session as a sequence of queries. The user's search intent is then formulated using query encoding and its corresponding query session encoding. Finally, in the decoding process, a recurrent neural decoder generates the refined query. Dehghani et al. [22] employed a seq-to-seq model with a term-level attention layer to discern the relationships between terms in the original query and the refined query. Wu et al. [89] used a memory network designed to effectively model user feedback in the context of information retrieval. Finally, Ahmed et al. [5] suggested incorporating historical (query, clicked documents) pairs to learn multitask of query suggestions and document ranking in tandem.

Supervised and semi-supervised methods, by and large, are trained on search logs, assuming that a user would gradually refine her query over successive attempts to find relevant content within a search session, which has been *in*validated by Chen et al. [19]; a user might search for multiple topics in one session, and hence, *ir*relevant queries would be learnt to be paired as $(q \rightarrow q^{\star})$. Moreover, search engines that are not on the web and built for customized applications may lack search logs, especially when a system is newly deployed, and even later on, the log rarely becomes as rich as that of web search engines [10].

Recently, we have observed new research efforts to produce standard benchmark datasets free of topic drifts that are specifically designed to train and evaluate the efficacy of supervised query refiners for information retrieval systems [8, 81, 96]. Among the first, Tamannaee et al. [81] proposed a configurable and reproducible pipeline to generate *gold*-standard datasets for a set of original queries by applying a host of more than twenty unsupervised query refiners, from simple lexical lemmatizers to complex pseudo-relevance-based methods. Then, those versions of the original query that improved the performance of a retrieval method were kept as refined queries. This way, both the original query and the refined versions were almost surely guaranteed to be in the same semantic context. Such a pipeline is, however, computationally costly for large-scale query sets due to its exhaustive application of refiners on each query. Also, despite many variations of an original query, the outcomes often yielded little to no refined query. To address scalability, Arabzadeh et al. [8] fed a query to pretrained doct5query [64] transformer and selected the generated sequence of tokens as a refined version should it increases bm25 retrieval performance based on map metric. However, Arabzadeh et al.'s work is case-specific, considerably less extensible, and heavily depends on doct5query; hence, it is incapable of accommodating different or new query sets, let alone no implementation is publicly available; only the final generated dataset is publicly released. Building upon Arabzadeh et al.'s work, Narayanan et al. [62] have developed an open-source reproducible pipeline to generate benchmark datasets from various domains and any choice of transformers. However, fine-tuning or aligning transformers can be computationally intensive and environmentally unsustainable due to the significant energy consumption of training large models on powerful hardware. To the best of our knowledge, no one has yet explored the synergistic impact of backtranslation in query refinement.

[2] A language family is a set of languages which share cultural roots and exhibit similarities in vocabulary and grammar [7].

Table 2: Notations used in this paper.

notation	description
r	an information retrieval method (retriever)
m	an information retrieval evaluation metric
q	an original query where $m_r(q, \mathcal{J}_q) < 1$
$\mathcal{J}_q$	the reference set of relevant documents (relevance judgment) for q
$m_r(q, \mathcal{J}_q)$	the performance of r to retrieve relevant documents for q in terms of m
q_l	a backtranslated query via language l for q
$q^\diamond$	a refined query for q, i.e., $m_r(q, \mathcal{J}_q) < m_r(q^\diamond, \mathcal{J}_q)$
$\mathcal{R}_{q,r,m}$	the set of all refined queries for q
$q^\star$	the best refined query for q where $\texttt{argmax}_{q^\diamond \in \mathcal{R}_{q,r,m}} m_r(q^\diamond, \mathcal{J}_q)$
$\bar{q}$	a hard (difficult) query, i.e., $\mathcal{R}_{\bar{q},r,m} = \emptyset$
Δ	the efficacy improvement of the metric m, i.e., $m_r(q^\star, \mathcal{J}_q) - m_r(q, \mathcal{J}_q)$

2.2 Natural Language Backtranslation

Backtranslation yields a new version of the sentence with different and diverse wordings while the meaning remains intact, and hence, has found immediate applications for a wide range of natural language processing tasks as a (1) data augmentation technique such as in machine translation [25, 48, 76], document classification [38], review analysis [36, 50], and question-answering [9], or (2) as a quality estimator in evaluating the quality of translations without human-translated references [3, 61, 99].

As an augmentation technique, Li et al. [48] and Haq et al. [85] employed backtranslation to generate synthetic parallel corpora in low-resource languages and to scale up the training set for neural machine translators. Ibrahim et al. [38] tackled the class imbalance in training sets for online offensive content detection. Hemmatizadeh et al. [36] tapped into backtranslation to empower the aspect-based sentiment classifiers and aspect detection methods with *latent* aspect detection. Bhaisaheb et al. [9] iteratively augmented a set of reasoning questions about data charts to leverage *compositional generalization*, i.e., producing *unseen* meaningful combinations of seen terms in sentences, and to improve generating analytical answers via sql programs using `codet5` [88].

For quality estimation, Moon et al. [61] and others [3, 99] use backtranslation as a semantic-level metric for multilingual two-way machine translation when no human-translated reference is available. The approach mimics end-users who assess the quality of an online multilingual translator by comparing the original sentence in a source language and the backtranslated sentence via a target language that they do not understand. Backtranslation as a quality metric outweighs reference-based metrics such as `blue`, which are limited to evaluating the surface-level lexical similarity.

Nonetheless, while backtranslation has been widely employed in nlp tasks, its effectiveness for query refinement in information retrieval has remained unclear, and we are the first to investigate it.

3 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 $q^\star$ to the *best* refined query, i.e., $q^\star = \texttt{argmax}_{q^\diamond \in \mathcal{R}_{q,r,m}} m_r(q^\diamond, \mathcal{J}_q)$. 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 $\mathcal{R}_{q,r,m} = q = q^\star$, and hence, query refinement is unnecessary.

4 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 workflow is shown in Figure 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.

4.1 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 [23, 32]. While languages share underlying commonalities referred to as linguistic *universals* due to the common neurobiological basis of the human brain [29], 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 [18]. 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, 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 *'making'* 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) [84], 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.

4.2 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:

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

where $m_r(., \mathcal{J}_q)$ 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 $R_{q,r,m}$ 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.

5 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.

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 [84] and bing [55].

5.1 Setup

5.1.1 Query Sets.

Our benchmark includes well-known query sets in english from different domains, namely dbpedia [11, 35] collection of wikipedia articles, robust04 [87] collection of news articles and US government publications, antique’s test collection [34] including open-domain non-factoid questions from Yahoo! Answers, gov2 [20] webpages of .gov web domain, and clueweb09b [21] 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 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.

5.1.2 Query Backtranslation.

We leverage Meta’s *‘no language left behind’* (nllb) [84][3], 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 [58] and is developed based on a conditional mixture of several transformers [77] that is trained on data tailored for low-resource languages. On the other extreme, we alternatively chose the bing translator[4], a cloud-based *closed*-source machine translation service offered by Microsoft [55, 56] 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.

[3] https://github.com/facebookresearch/fairseq/tree/nllb
[4] https://www.bing.com/Translator

Table 3: 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 $\lvert q\rvert$	$\lvert\mathcal{J}_.\rvert$	$\#q : \mathcal{J}_q = \emptyset$	map	ndcg	mrr	map	ndcg	mrr	map	ndcg	mrr	map	ndcg	mrr
dbpedia [11, 35]	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	12	10	258
robust04 [87]	news	250	528,155	2.76	311,410	1	0.199	0.368	0.667	0.201	0.373	0.681	1	1	138	1	1	143
antique [34]	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	1	0	123
gov2 [20]	*.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	1	1	89
clueweb09b [21]	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	2	2	55

Table 4: 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.

		dbpedia						robust04						antique						gov2						clueweb09b					
family	language	$\lvert q_l\rvert - \lvert q\rvert$		declutr [30]		rouge-l		$\lvert q_l\rvert - \lvert q\rvert$		declutr [30]		rouge-l		$\lvert q_l\rvert - \lvert q\rvert$		declutr [30]		rouge-l		$\lvert q_l\rvert - \lvert q\rvert$		declutr [30]		rouge-l		$\lvert q_l\rvert - \lvert q\rvert$		declutr [30]		rouge-l	
		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
indo-european	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
	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

We translate queries from english into 10 languages from 7 language families, including malay, swahili, and tamil as low-resource languages. Table 4 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-l [51] and declutr by Giorgi et al. [30]. Backtranslation from english to itself has been performed for unit test purposes where all the results for declutr and rouge-l 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-l and declutr scores. In terms of translation quality, while rouge-l 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-l 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**.

5.1.3 *Gold-standard Dataset Generation.*

We have applied two *sparse* information retrieval methods, namely bm25 [69] and qld [67], using pyserini [52] to retrieve relevant content for the original queries as well as the backtranslated versions. We acknowledge dense information retrieval methods like colbert [41] 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 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.

```
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: 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.

We evaluate the retrieval performances based on three metrics, i.e., map, mrr and ndcg, using trec_eval [66]. Those backtranslated versions that increased a metric value form a gold-standard dataset. In total, we generate a family of {dbpedia, robust04, antique, gov2, clueweb09b} × {bm25, qld} × {map, mrr, ndcg} = 30 gold-standard datasets. Figure 2 shows the file structure of the gold-standard dataset in robust04.bm25.map.tsv.

5.1.4 *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 [28], 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 [73],
- semantic refiners, which use an external linguistic knowledge-base including thesaurus [78], wordnet [65], and conceptnet [1], to extract related terms to the original query's terms,

Table 5: Efficacy of backtranslated queries in query refinement. $\#q$ shows the number of original queries that need refinement, while $\#q^\star$ and % represent the *best* refined queries' count and percentage, respectively, and Δ denotes the average metric improvements.

		bm25				qld			
		$\#q$	$\#q^\star$	%	Δ	$\#q$	$\#q^\star$	%	Δ
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

- sense-disambiguation [82], 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 [2] and word2vec [57] to find the most similar terms to the query terms,
- anchor [43], 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 [6], which uses the embeddings trained on wikipedia's hierarchical categories [49] 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 [72], 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 [16], docluster [45], and conceptluster [63], where a graph clustering method like Louvain [12] 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 [98], which employs bert's contextualized word embeddings of terms in the top-k retrieved documents.

(2) To evaluate whether the expanded gold-standard datasets in 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) [68], 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 [86]. 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 top-$k = 10$ random sampling during inference. We use 70% of ($q \rightarrow q^\star$) 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.

Table 6: Results of t5 [68] 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	**0.336**	0.279	0.496	**0.505**	0.404	0.768	**0.791**
robust04.bm25.map.tsv	0.167	0.277	**0.286**	0.323	0.464	**0.475**	0.605	0.824	**0.841**
antique.bm25.map.tsv	0.227	0.488	**0.494**	0.342	0.591	**0.597**	0.634	0.972	**0.979**
gov2.bm25.map.tsv	0.134	0.225	**0.228**	0.276	0.390	**0.393**	0.677	0.848	**0.869**

5.2 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 5, 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 backtraslations, which can be used to augment training sets for supervised query refiners. Moreover, from Table 6, 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 has not seen any training pairs.

To respond **RQ2**, we compared query backtranslation with global and local unsupervised refiners [81]. In Table 7, 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 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 7: Distribution of refined queries across refinement methods, including query backtranslation, local and global unsupervised refiners in terms of map; $\#q^\star$ 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	
		$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%	$\#q^\star$	%
	backtranslation [ours]	65	13.92	47	**18.88**	17	8.50	17	**11.41**	12	6.06	72	**15.42**	47	**18.88**	29	14.50	14	9.40	9	4.55
global	tagme [28]	70	**14.99**	19	7.63	19	9.50	13	8.72	22	**11.11**	62	13.28	20	8.03	17	8.50	12	8.05	19	**9.60**
	thesaurus [78]	34	7.28	0	0.00	114	**57.00**	0	0.00	1	0.51	38	8.14	0	0.00	102	**51.00**	0	0.00	1	0.51
	wiki [6]	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	7.58
	anchor [43]	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 [1]	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 [2]	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 [82]	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 [57]	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 [65]	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 [73]	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 [73]	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 [73]	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 [73]	2	0.43	2	0.80	11	5.50	0	0.00	0	0.00	1	0.21	1	0.40	1	0.50	0	0.00	0	0.00
	stem.remover [73]	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 [73]	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 [73]	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
local	relevance-feedback [72]	36	7.71	47	**18.88**	6	3.00	15	10.07	16	8.08	25	5.35	39	15.66	5	2.50	12	8.05	19	**9.60**
	termluster [16]	0	0.00	0	0.00	0	0.00	17	**11.41**	3	1.52	0	0.00	0	0.00	0	0.00	16	10.74	6	3.03
	rm3 [17]	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	**13.42**	2	1.01
	bertqe [98]	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
	conceptluster [63]	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
	docluster [45]	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 ($\#\bar{q}$)		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 unrefind 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 3, 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.

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 8 and Figure 5 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 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.

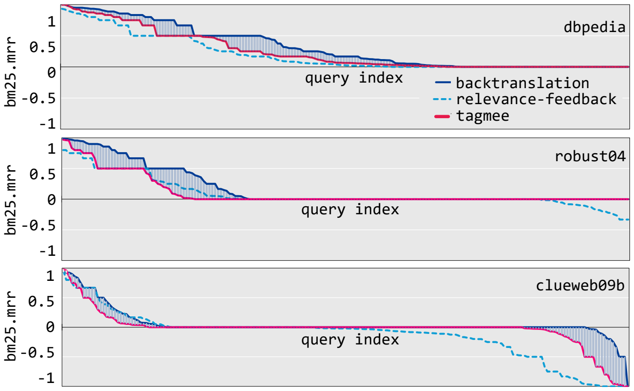

Figure 3: Distribution of Δmrr across original queries for backtranslation vs. relevance-feedback, and tagme.

With respect to **RQ4**, from Tables 5 and 8, 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 5, 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 4 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

Table 8: Efficacy of query backtranslation across languages; % shows the percentage of queries matched with a refined query, and Δ shows the average metric improvements. Bold and underlined numbers are *row-wise* highest and second-highest, respectively.

			indo-european								austronesian		dravidian		bantu		sino-tibetan		koreanic		afro-asiatic	
			farsi		french		german		russian		malay		tamil		swahili		chinese		korean		arabic	
		$\#q$	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%	Δ
bm25.map	dbpedia	460	12.17	+0.07	**17.17**	+0.09	12.61	+0.10	16.30	+0.12	16.96	+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	16.47	+0.05	**16.87**	+0.06	12.85	+0.09
	antique	200	04.50	+0.07	10.50	+0.06	10.00	+0.05	**11.00**	+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	**10.74**	+0.07	08.72	+0.05	08.05	+0.03	06.71	+0.05	10.07	+0.04	10.07	+0.05	06.71	+0.03
	clueweb09b	198	02.53	+0.07	04.04	+0.02	02.53	+0.04	02.53	+0.04	02.53	+0.04	02.02	+0.04	02.02	+0.01	**05.56**	+0.01	03.03	+0.01	02.53	+0.05
bm25.ndcg	dbpedia	461	13.45	+0.10	**17.57**	+0.11	13.23	+0.12	16.05	+0.14	17.14	+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	**17.27**	+0.08	14.46	+0.08	12.45	+0.11	12.85	+0.09	16.47	+0.08	16.06	+0.10	12.05	+0.12
	antique	200	07.50	+0.09	12.00	+0.07	10.00	+0.05	**12.50**	+0.07	12.00	+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	10.07	+0.07	09.40	+0.07	07.38	+0.03	07.38	+0.04	**10.74**	+0.04	**10.74**	+0.06	06.71	+0.04
	clueweb09b	198	02.53	+0.07	03.54	+0.06	01.52	+0.10	2.53	+0.05	01.52	+0.09	02.53	+0.05	0.51	+0.05	**05.05**	+0.03	03.03	+0.03	02.53	+0.06
bm25.mrr	dbpedia	255	18.43	+0.28	23.53	+0.33	18.82	+0.38	22.35	+0.34	**23.92**	+0.35	21.18	+0.34	20.00	+0.32	23.53	+0.40	19.22	+0.35	20.00	+0.38
	robust04	112	16.96	+0.44	20.54	+0.32	20.54	+0.42	**24.11**	+0.44	19.64	+0.47	**24.11**	+0.43	18.75	+0.37	23.21	+0.40	23.21	+0.32	21.43	+0.38
	antique	037	21.62	+0.40	24.32	+0.50	24.32	+0.35	**27.03**	+0.52	**27.03**	+0.53	24.32	+0.35	13.51	+0.40	**27.03**	+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	**28.07**	+0.45	22.81	+0.37	15.79	+0.42
	clueweb09b	145	04.14	+0.53	04.83	+0.36	04.14	+0.36	04.83	+0.41	06.21	+0.38	06.21	+0.39	02.76	+0.43	**08.28**	+0.32	**08.28**	+0.36	05.52	+0.42

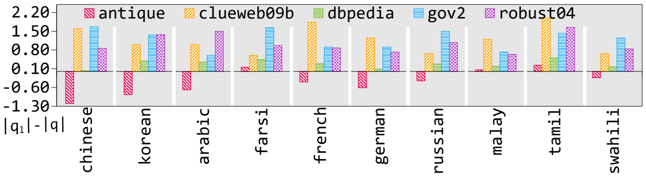

Figure 4: The length difference between refined query via backtranslation vs. original query.

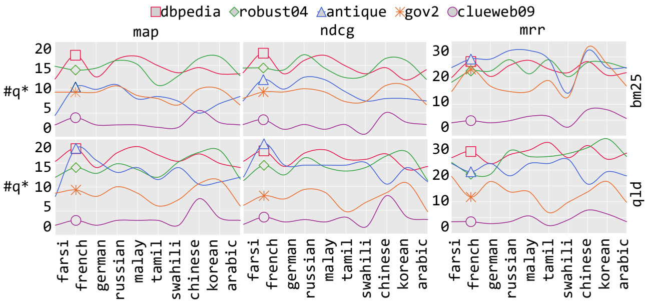

Figure 5: The language spectrum to illustrate the influence of language across each query set based on the number of *best* refined query obtained by each language.

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 cluweb09b compared to long queries from a corpus with a limited span of topics like dbpedia.

To answer **RQ5**, i.e., the efficacy of query backtranslation across different translators, Table 9 shows a comparison between our choice of translator from Meta's nllb [84] and an alternative closed-source translator from Microsoft bing [55]. As seen, the application of nllb notably yields more refined queries, and bing performed poorly. Meanwhile, looking at their translation qualities in Table 4, 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-l compared to bing. Table 9 and Table 4 together underline that a translator that accurately but with more diverse paraphrases would yield more refined queries.

Table 9: Meta's nllb vs. Microsoft's bing in query refinement.

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

6 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] [n. d.]. ConceptNet. http://conceptnet.io/.

[2] [n. d.]. GloVe. https://nlp.stanford.edu/projects/glove/.

[3] Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality Estimation via Backtranslation at the WMT 2022 Quality Estimation Task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, Philipp Koehn, Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (Eds.). Association for Computational Linguistics, 593–596. https://aclanthology.org/2022.wmt-1.54

[4] Sanjay Agrawal, Srujana Merugu, and Vivek Sembium. 2023. 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*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 4488–4494. https://doi.org/10.1145/3583780.3615474

[5] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 385–394.

[6] Bashar Al-Shboul and Sung-Hyon Myaeng. 2014. Wikipedia-based query phrase expansion in patent class search. *Information retrieval* 17 (2014), 430–451.

[7] Stephen R. Anderson. 2012. 10How many languages are there in the world? In *Languages: A Very Short Introduction.* Oxford University Press. https://doi.org/10.1093/actrade/9780199590599.003.0002

[8] Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. Matches Made in Heaven: Toolkit and Large-Scale Datasets for Supervised Query Reformulation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* 4417–4425.

[9] Shabbirhussain Bhaisaheb, Shubham Paliwal, Rajaswa Patil, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2023. 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*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2501–2515. https://doi.org/10.18653/v1/2023.findings-eacl.189

[10] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. 2011. 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*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 795–804. https://doi.org/10.1145/2009916.2010023

[11] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of web semantics* 7, 3 (2009), 154–165.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[13] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. 1995. Automatic query expansion using SMART: TREC 3. *NIST special publication sp* (1995), 69–69.

[14] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* 243–250.

[15] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. 2009. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web.* 191–200.

[16] Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)* 19, 1 (2001), 1–27.

[17] Marc-Allen Cartright, James Allan, Victor Lavrenko, and Andrew McGregor. 2010. Fast query expansion using approximations of relevance models. In *Proceedings of the 19th ACM international conference on Information and knowledge management.* 1573–1576.

[18] Damir Cavar, Ludovic Mompelat, and Muhammad Abdo. 2024. 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*, Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, and Ekaterina Vylomova (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 46–54. https://aclanthology.org/2024.sigtyp-1.6

[19] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. 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*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 743–755. https://doi.org/10.1145/3442381.3450127

[20] Charles LA Clarke, Falk Scholer, and Ian Soboroff. 2005. The TREC 2005 Terabyte Track.. In *TREC.*

[21] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *TREC.*

[22] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 1747–1756.

[23] Stefano Demichelis and Jörgen W Weibull. 2008. Language, meaning, and games: A model of communication, coordination, and evolution. *American Economic Review* 98, 4 (2008), 1292–1311.

[24] Doug Downey, Susan Dumais, and Eric Horvitz. 2007. Heads and tails: studies of web search with common and rare queries. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* 847–848.

[25] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. 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.* Association for Computational Linguistics, Online, 704–717. https://doi.org/10.18653/v1/2021.naacl-main.57

[26] Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq R. Joty, Dragomir R. Radev, and Yashar Mehdad. 2021. 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*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 704–717. https://doi.org/10.18653/v1/2021.naacl-main.57

[27] Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. 2021. 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*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 553–564. https://doi.org/10.1145/3442381.3449789

[28] Paolo Ferragina and Ugo Scaiella. 2010. 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* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1625–1628. https://doi.org/10.1145/1871437.1871689

[29] Angela D Friederici. 2017. *Language in our brain: The origins of a uniquely human capacity.* MIT Press.

[30] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. 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*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 879–895. https://doi.org/10.18653/v1/2021.acl-long.72

[31] Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. 2021. 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, 7601–7609. https://doi.org/10.1609/AAAI.V35I9.16930

[32] Joan Kelly Hall. 2013. *Teaching and researching: Language and culture.* Routledge.

[33] Fred X Han, Di Niu, Haolan Chen, Kunfeng Lai, Yancheng He, and Yu Xu. 2019. A deep generative approach to search extrapolation and recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 1771–1779.

[34] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. 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, 166–173.

[35] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. 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*. 1265–1268.

[36] Farinam Hemmatizadeh, Chrsitine Wong, Alice Yu, and Hossein Fani. 2023. 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. https://doi.org/10.1145/3583780.3615205

[37] Xing Hu, Ling Liang, Xiaobing Chen, Lei Deng, Yu Ji, Yufei Ding, Zidong Du, Qi Guo, Timothy Sherwood, and Yuan Xie. 2022. A Systematic View of Model Leakage Risks in Deep Neural Network Systems. *IEEE Trans. Computers* 71, 12 (2022), 3254–3267. https://doi.org/10.1109/TC.2022.3148235

[38] Mai Ibrahim, Marwan Torki, and Nagwa M. El-Makky. 2020. AlexU-BackTranslation-TL at SemEval-2020 Task 12: Improving Offensive Language Detection Using Data Augmentation and Transfer Learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, Aurélie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, 1881–1890. https://doi.org/10.18653/v1/2020.semeval-1.248

[39] Siti Nurkhadijah Aishah Ibrahim, Ali Selamat, and Mohd Hafiz Selamat. 2009. Query optimization in relevance feedback using hybrid GA-PSO for effective web information retrieval. In *2009 Third Asia International Conference on Modelling & Simulation*. IEEE, 91–96.

[40] K Sparck Jones and EO Barber. 1971. What makes an automatic keyword classification effective? *Journal of the American Society for Information Science* 22, 3 (1971), 166–175.

[41] Omar Khattab and Matei Zaharia. 2020. 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*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[42] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 67–72. https://www.aclweb.org/anthology/P17-4012

[43] Reiner Kraft and Jason Zien. 2004. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*. 666–674.

[44] Oh-Woog Kwon, Myoung-Cheol Kim, and Key-Sun Choi. 1994. Query expansion using domain-adapted, weighted thesaurus in an extended Boolean model. In *Proceedings of the third international conference on Information and knowledge management*. 140–146.

[45] Kyung Soon Lee, W Bruce Croft, and James Allan. 2008. 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*. 235–242.

[46] Ruirui Li, Liangda Li, Xian Wu, Yunhong Zhou, and Wei Wang. 2019. Click feedback-aware query recommendation using adversarial examples. In *The World Wide Web Conference*. 2978–2984.

[47] Yu Li, Xiao Li, Yating Yang, and Rui Dong. 2020. A Diverse Data Augmentation Strategy for Low-Resource Neural Machine Translation. *Inf.* 11, 5 (2020), 255. https://doi.org/10.3390/info11050255

[48] Yu Li, Xiao Li, Yating Yang, and Rui Dong. 2020. A diverse data augmentation strategy for low-resource neural machine translation. *Information* 11, 5 (2020), 255.

[49] Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia P. Sycara. 2016. 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*, Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad (Eds.). ACL, 2678–2688. https://www.aclweb.org/anthology/C16-1252/

[50] Tomas Liesting, Flavius Frasincar, and Maria Mihaela Truşcă. 2021. Data Augmentation in a Hybrid Approach for Aspect-Based Sentiment Analysis. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (Virtual Event, Republic of Korea) *(SAC '21)*. Association for Computing Machinery, New York, NY, USA, 828–835. https://doi.org/10.1145/3412841.3441958

[51] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[52] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. 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)*. 2356–2362.

[53] Saurav Manchanda, Mohit Sharma, and George Karypis. 2019. Intent term selection and refinement in e-commerce queries. *arXiv preprint arXiv:1908.08564* (2019).

[54] Xiangke Mao, Shaobin Huang, Rongsheng Li, and Linshan Shen. 2020. Automatic keywords extraction based on co-occurrence and semantic relationships between words. *IEEE Access* 8 (2020), 117528–117538.

[55] Microsoft. [n. d.]. Microsoft Translator GitHub Repository. https://github.com/MicrosoftTranslator. Accessed: [Insert Date].

[56] Microsoft. 2023. *Azure AI Custom Translator Neural Dictionary Delivering Higher Terminology Translation Quality*. Microsoft. https://www.microsoft.com/en-us/translator/blog/2023/12/06/azure-ai-custom-translator-neural-dictionary-delivering-higher-terminology-translation-quality/

[57] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[58] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 220–229. https://doi.org/10.1145/3287560.3287596

[59] Hans Moen, Laura-Maria Peltonen, Henry Suhonen, Hanna-Maria Matinolli, Riitta Mieronkoski, Kirsi Telen, Kirsi Terho, Tapio Salakoski, and Sanna Salanterä. 2019. An Unsupervised Query Rewriting Approach Using N-gram Co-occurrence Statistics to Find Similar Phrases in Large Text Corpora. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press, Turku, Finland, 131–139. https://aclanthology.org/W19-6114

[60] Akash Kumar Mohankumar, Nikit Begwani, and Amit Singh. 2021. 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*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 3423–3431. https://doi.org/10.1145/3447548.3467202

[61] Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting Round-trip Translation for Quality Estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, Mikel L. Forcada, André Martins, Helena Moniz, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof Arenas, Mary Nurminen, Lena Marg, Sara Fumega, Bruno Martins, Fernando Batista, Luísa Coheur, Carla Parra Escartín, and Isabel Trancoso (Eds.). European Association for Machine Translation, 91–104. https://aclanthology.org/2020.eamt-1.11/

[62] Yogeswar Lakshmi Narayanan and Hossein Fani. 2023. 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. https://doi.org/10.1145/3583780.3615129

[63] Apostol Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th ACM international conference on Multimedia*. 991–1000.

[64] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6 (2019), 2.

[65] Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. 2014. Improving query expansion using WordNet. *Journal of the Association for Information Science and Technology* 65, 12 (2014), 2469–2478.

[66] Joao Palotti, Harrisen Scells, and Guido Zuccon. 2019. TrecTools: an open-source Python library for Information Retrieval practitioners involved in TREC-like campaigns *(SIGIR'19)*. ACM.

[67] Jay M Ponte and W Bruce Croft. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 202–208.

[68] Colin Raffel and et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.

[69] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[70] J. J. Rocchio. 1971. Relevance feedback in information retrieval. In *The Smart retrieval system - experiments in automatic document processing*, G. Salton (Ed.). Englewood Cliffs, NJ: Prentice-Hall, 313–323.

[71] Haggai Roitman, Ella Rabinovich, and Oren Sar Shalom. 2018. 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*, Dongwon Lee, Nishanth Sastry, and Ingmar Weber (Eds.). ACM, 33–37. https://doi.org/10.1145/3209542.3209567

[72] Gerard Salton. 1971. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc.

[73] Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for*

Computational Linguistics 4 (2016), 287–300.

[74] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM* 63, 12 (2020), 54–63.

[75] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. https://doi.org/10.18653/v1/p16-1009

[76] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 86–96. https://doi.org/10.18653/v1/P16-1009

[77] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. 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. https://openreview.net/forum?id=B1ckMDqlg

[78] Ali Asghar Shiri. 2003. End-user interaction with thesaurus-enhanced search interfaces, an evaluation of search term selection for query expansion. (2003).

[79] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[80] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management.* 553–562.

[81] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. Reque: a configurable workflow and dataset collection for query refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 3165–3172.

[82] Liling Tan. 2014. Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software].

[83] Tao Tao and ChengXiang Zhai. 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* 162–169.

[84] NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling human-centered machine translation. *línea]. Disponible en: https://github. com/facebookresearch/fairseq/tree/nllb* (2022).

[85] Sami ul Haq, Sadaf Abdul-Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document Level NMT of Low-Resource Languages with Backtranslation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (Eds.). Association for Computational Linguistics, 442–446. https://aclanthology.org/2020.wmt-1.53/

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 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*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[87] Ellen Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. https://doi.org/10.6028/NIST.SP.500-261

[88] Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8696–8708. https://doi.org/10.18653/v1/2021.emnlp-main.685

[89] Bin Wu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Query suggestion with feedback memory network. In *Proceedings of the 2018 World Wide Web Conference.* 1563–1571.

[90] Yonghao Wu, Zheng Li, Jie M. Zhang, and Yong Liu. 2023. ConDefects: A New Dataset to Address the Data Leakage Concern for LLM-based Fault Localization and Program Repair. *CoRR* abs/2310.16253 (2023). https://doi.org/10.48550/ARXIV.2310.16253 arXiv:2310.16253

[91] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144 (2016). arXiv:1609.08144 http://arxiv.org/abs/1609.08144

[92] Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. 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, 9298–9305. https://ojs.aaai.org/index.php/AAAI/article/view/6469

[93] Jinxi Xu and W Bruce Croft. 2017. Quary expansion using local and global document analysis. In *Acm sigir forum*, Vol. 51. ACM New York, NY, USA, 168–175.

[94] Dayu Yang, Yue Zhang, and Hui Fang. 2022. An Exploration Study of Mixed-initiative Query Reformulation in Conversational Passage Retrieval. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/udel_fang.C.pdf

[95] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. [n. d.]. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541* ([n. d.]).

[96] George Zerveas, Ruochen Zhang, Leila Kim, and Carsten Eickhoff. 2020. Brown University at TREC Deep Learning 2019. *CoRR* abs/2009.04016 (2020). arXiv:2009.04016 https://arxiv.org/abs/2009.04016

[97] Xiaojuan Zhang. 2022. Improving personalised query reformulation with embeddings. *Journal of Information Science* 48, 4 (2022), 503–523.

[98] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 4718–4728. https://www.aclweb.org/anthology/2020.findings-emnlp.424

[99] Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking Round-Trip Translation for Machine Translation Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 319–337. https://doi.org/10.18653/v1/2023.findings-acl.22

[100] Xiaochen Zuo, Zhicheng Dou, and Ji-Rong Wen. 2022. Improving session search by modeling multi-granularity historical query change. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining.* 1534–1542.