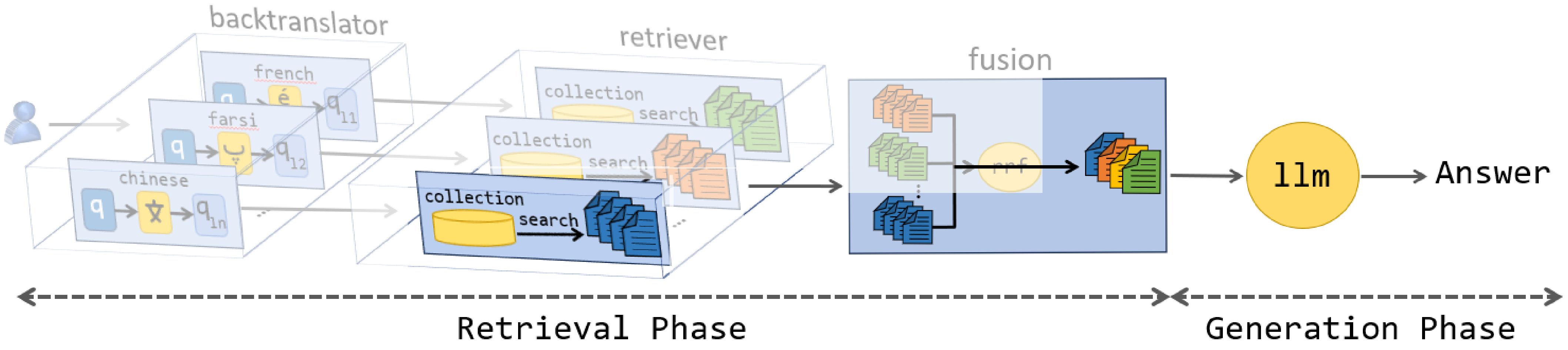
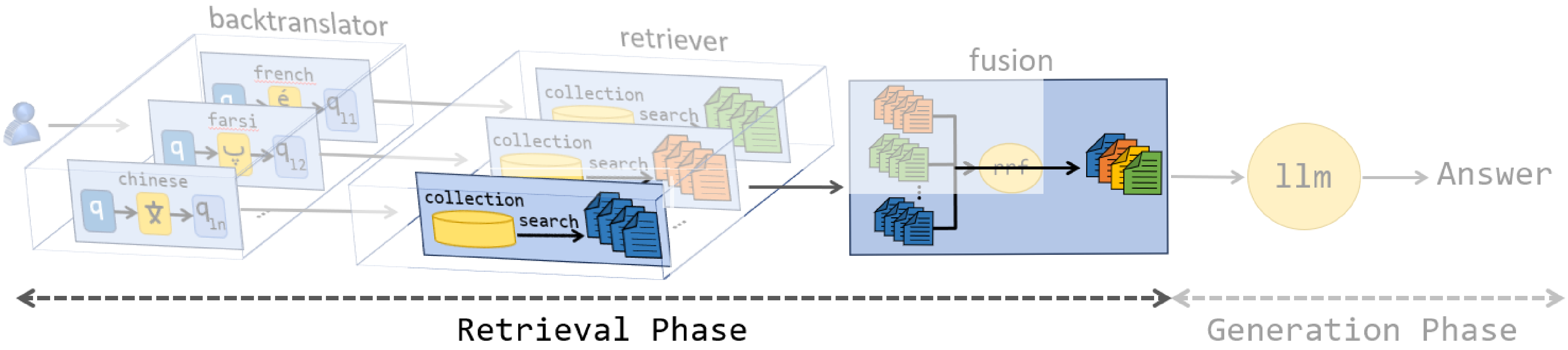
Enhancing RAG's Retrieval via Query Backtranslations

WISE 2024

Retrieval-augmented generation (rag)



Retrieval-augmented generation (rag)



Retrieval-augmented generation (rag)

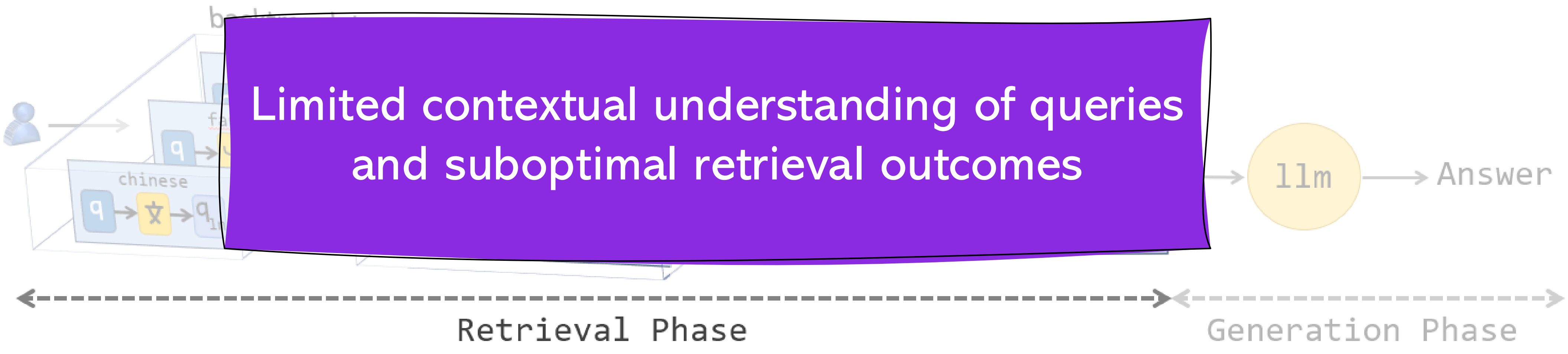
- Improving the accuracy of the document list has been shown to enhance the subsequent generation stage (NeurIPS '20, NAACL '22, CoRR '21)
- Retrieval component

→ A prebuilt retrieval model: 1) commercial search engines, 2)neural ranking models, 3)term matching retrieval models

→ Developing a custom retrieval model

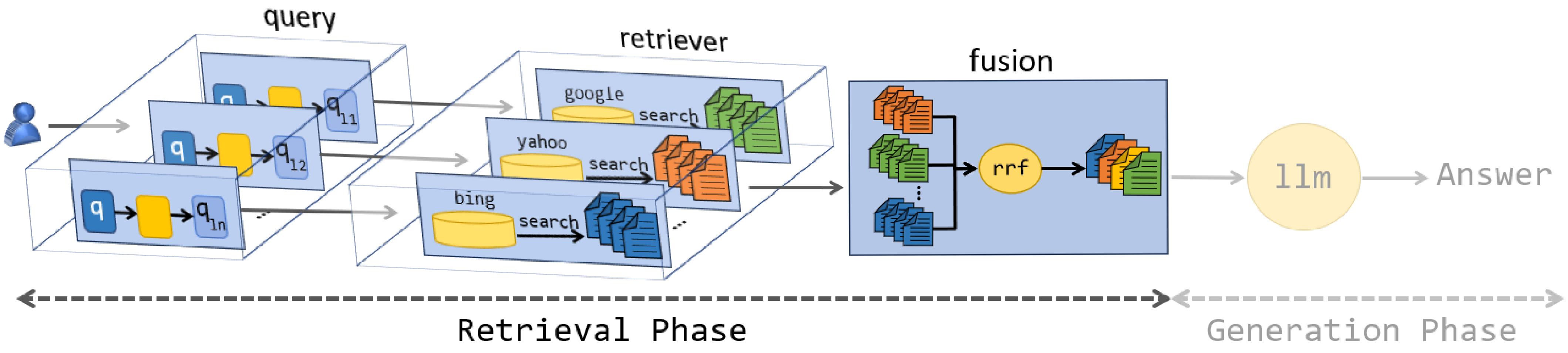
Retrieval-augmented generation for knowledge-intensive NLP tasks, NeurIPS , 2020
Re2g: Retrieve, rerank, generate, NAACL, 2022
Webgpt: Browser-assisted question-answering with human feedback, CoRR , 2021

Retrieval-augmented generation (rag)

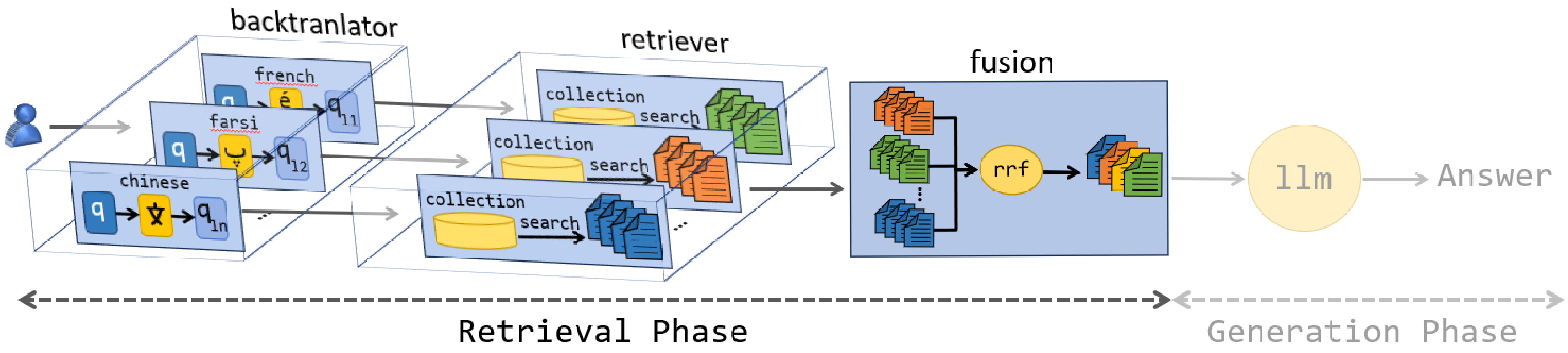


Limited contextual understanding of queries and suboptimal retrieval outcomes

Rag-fusion


query
retriever
fusion
q
q
q
google
search
yahoo
search
bing
search
rrf
llm
Answer
Retrieval Phase
Generation Phase

Backtranslation



Methodology

Backtranslation

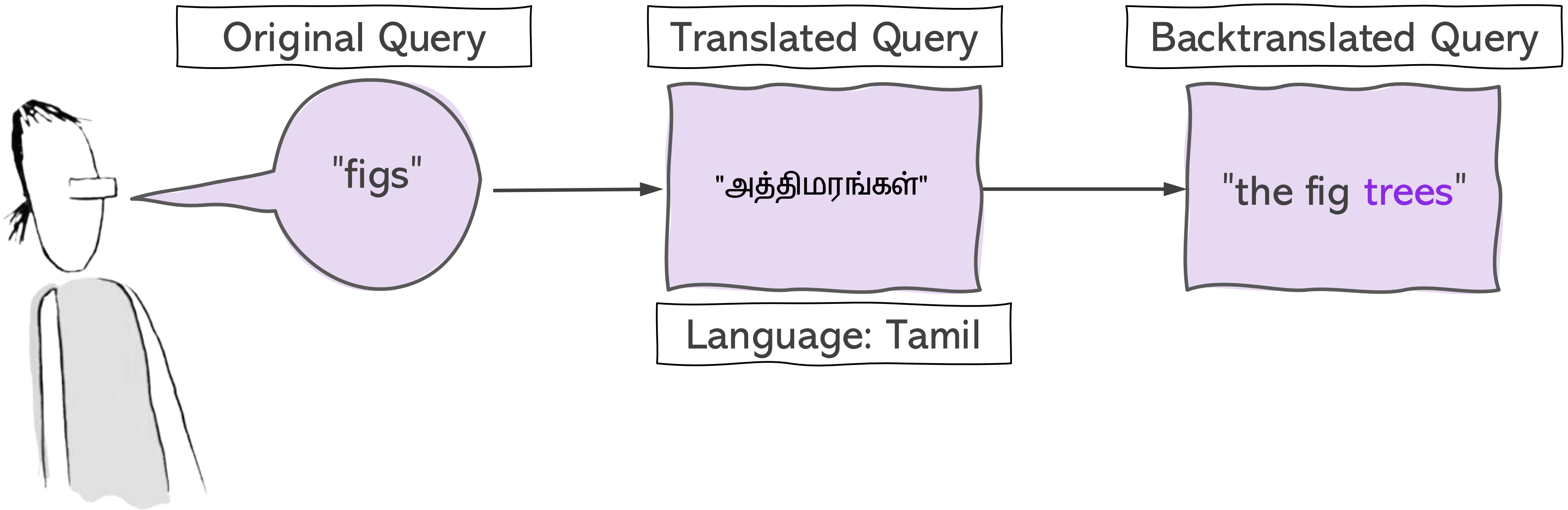
The basic idea of Backtranslation is to translate a sentence or a text from one language to another and then translate it back to the original language.

Original Query	“italian nobel prize winners"
Translated Query	”برندهای جایزه نوبل ایتالیایی“
Backtranslated Query	“italian nobel laureates”

Backtranslation Benefits

- **Revealing latent aspects**

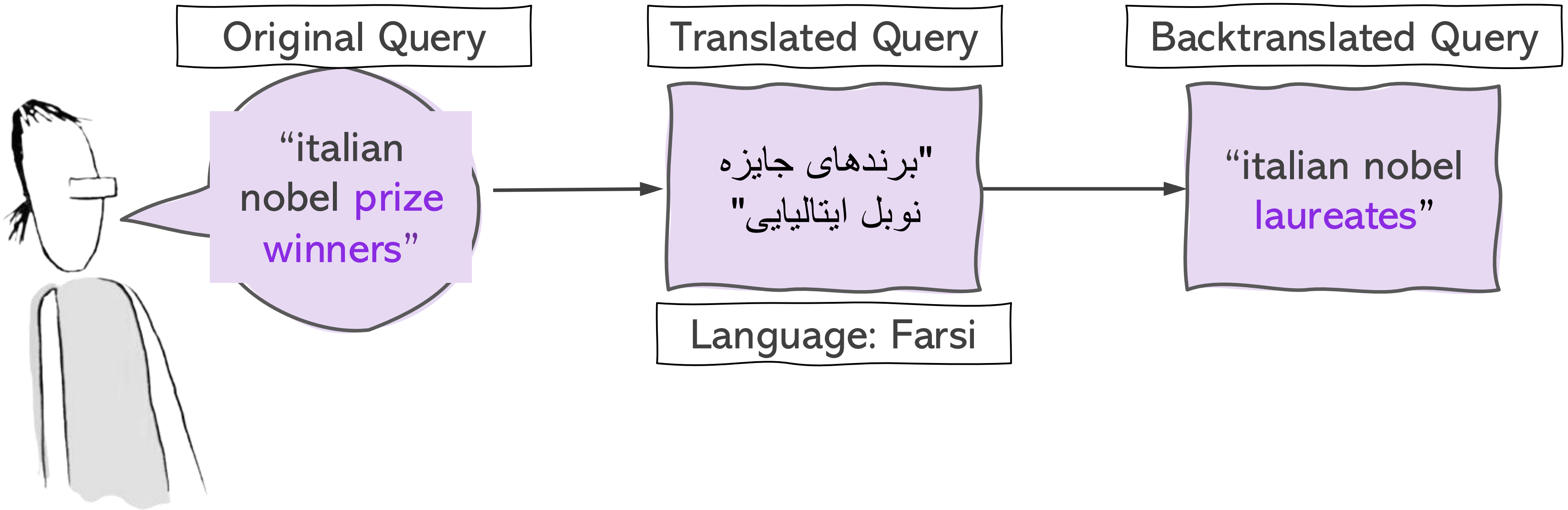
Backtranslation can uncover terms or entities that have not been explicitly mentioned in a query.



Backtranslation Benefits

- **Context-aware synonymous aspects**

  Backtranslation can augment context-aware <u>synonymous terms.</u>

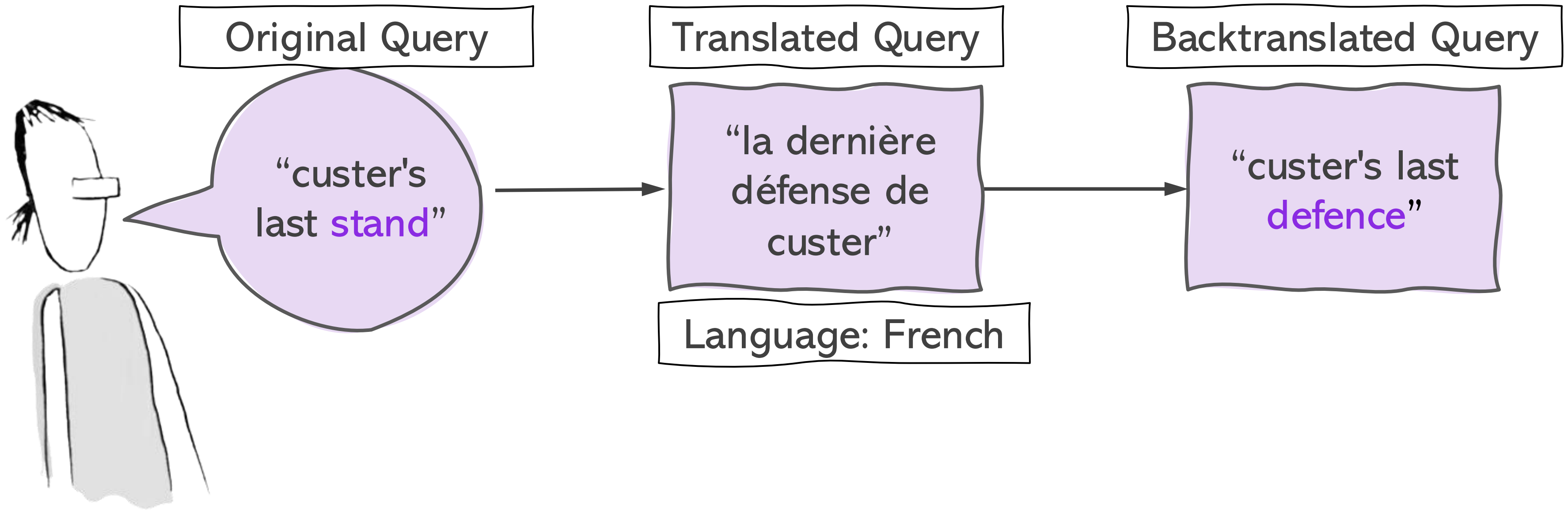


Methodology

Backtranslation Benefits

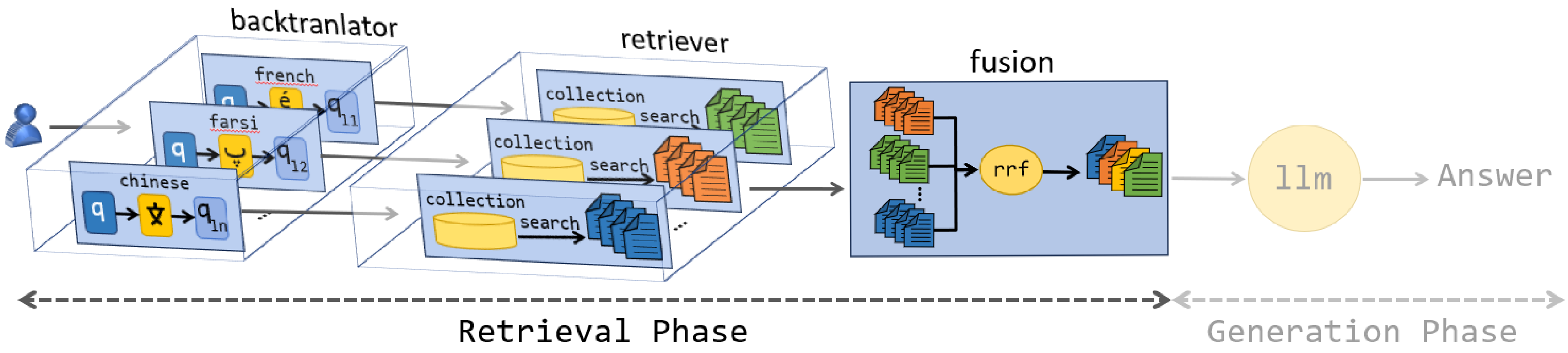
- **Clarifying the semantic disambiguation**

Backtranslation can disambiguate polysemous terms and collocations.



Methodology

Main Flow



Rag | Rag-fusion | Backtranslation | Methodology | Experiments | Results | Conclusion

Backtranslation

We have made use of the Meta (Facebook)'s **No Language Left Behind (NLLB)** neural machine translator for backtranslation task (EMNLP '22)

Why NLLB?

- Open-source machine translator
- Providing high-quality translations between 200 languages

No Language Left Behind: Scaling Human-Centered Machine Translation, NLLB Team, EMNLP, 2022

Methodology

Rag-based retrieval

we apply reciprocal rank fusion (rrf) (SIGIR 2009):

$$\mathrm{rrf}(d \in \mathcal{D}_q^*) = \sum_{\mathcal{D}_{q_l \in q_L}} \frac{1}{k + rank(d)}$$

Why rrf?

- We select reciprocal rank fusion because while highly ranked documents hold greater significance, the importance of lower-ranked ones should also be regarded.

Reciprocal rank fusion outperforms condorcet and individual rank learning methods, SIGIR, 2009

Experiments

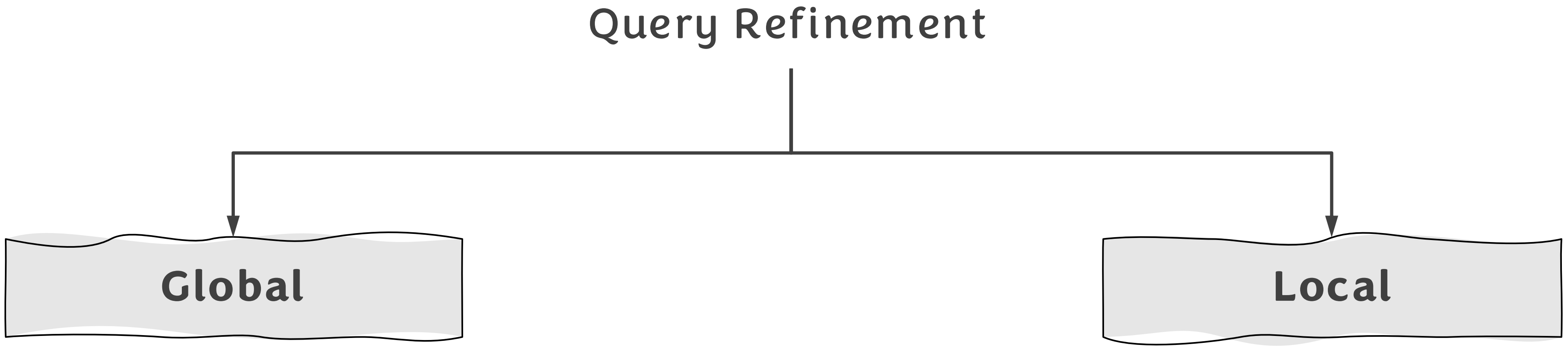
Datasets

Dataset	Description	#q	avg|q|
DBpedia	Wikipedia articles	467	5.37
Robust04	News articles and US government publications	250	2.76
ANTIQUE	Non-factoid question-answering by real users in Yahoo! Answers	200	9.34
GOV2	Substantial portion of webpages	150	3.13
ClueWeb09	Substantial portion of webpages	200	2.45

Dbpedia-a crystallization point for the web of data, Bizer et al, SSRN, 2009.
Overview of the TREC 2004 Robust Retrieval Track, Voorhees at al., NIST, 2005.
ANTIQUE: A non-factoid question answering benchmark, Hashemi et al., ECIR, 2020.
The TREC 2005 Terabyte Track, Clarke et al., TREC, 2005.
Overview of the TREC 2009 Web Track, Charles et al., TREC, 2009.

Experiments

Baseline



Baseline-global

Tagme	replaces the original query's terms with the title of their wikipedia articles
stemmers	utilize various lexical, syntactic, and semantic aspects of query to reduce the terms to their roots
semantic refiners	use an external linguistic knowledge-base
sense-disambiguation	resolves the ambiguity of polysemous terms in the original query based on the surrounding terms

Tagme: On-the-fly annotation of short text fragments, Ferragina et al, CIKM, 2010.
Overview of the TREC 2004 Robust Retrieval Track, Voorhees at al., NIST, 2005.
ANTIQUE: A non-factoid question answering benchmark, Hashemi et al., ECIR, 2020.
The TREC 2005 Terabyte Track, Clarke et al., TREC, 2005.

Baseline-global

embedding-based methods	use pre-trained term embeddings to find the most similar terms to the query terms
anchor	similar to embedding methods where the embeddings trained on wikipedia *anchors* texts
wiki	uses the embeddings trained on wikipedia's hierarchical categories
backtranslation	a query is translated from its original language to a set of target languages

Baseline-local

relevance-feedback	important terms from the top-k retrieved documents are added to the original query
clustering techniques	a graph clustering method ensures that each cluster consists of frequently co-occurring terms
bertqe	employs bert's contextualized word embeddings of terms in the top-k retrieved documents.

Experiments

Languages

English's language family

Source Language

Family	Language
Indo-European	Farsi
	French
	German
	Russian
Austronesian	Malay
Dravidian	Tamil
Bantu	Swahili
Sino-Tibetan	Chinese
Koreanic	Korean
Afro-Asiatic	Arabic

Target Language

Other Translators

For comparing quality of translators, we have selected **Bing Translator** neural machine translator for backtranslation task

Why Bing?

- Closed-source machine translator
- Providing high-quality translations between 128 languages

Azure AI Custom Translator Neural Dictionary Delivering Higher Terminology Translation Quality, Microsoft, EMNLP, 2023
www.microsoft.com/en-us/translator

Results

RQ1

		bm25.map									
		dbpedia		robust04		antique		gov2		clueweb09	
	reformulation method	#q**	%	#q**	%	#q**	%	#q**	%	#q**	%
rrf	rrf.all	**52**	11.13	33	13.25	17	8.50	**56**	37.58	**41**	20.81
	rrf.global	44	9.42	18	7.23	18	9.00	7	4.70	25	12.69
	rrf.local	37	7.92	12	4.82	38	19.00	18	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
	tagmee	49	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	**72**	36.00	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	relevance-feedback	16	3.43	**35**	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
	q	15	3.21	7	2.81	2	1.00	1	0.67	25	12.69
	sum	467	100	249	100	200	100	149	100	198	100

How does fusion perform across different query reformulation methods?

Rag | Rag-fusion | Backtranslation | Methodology | Experiments | Results | Conclusion

Results

RQ1

		bm25.map									
		dbpedia		robust04		antique		gov2		clueweb09	
	reformulation method	$\#q^{**}$	%	$\#q^{**}$	%	$\#q^{**}$	%	$\#q^{**}$	%	$\#q^{**}$	%
rrf	rrf.all	**52**	11.13	33	13.25	17	8.50	**56**	37.58	**41**	20.81
	rrf.global	44	9.42	18	7.23	18	9.00	7	4.70	25	12.69
	rrf.local	37	7.92	12	4.82	38	19.00	18	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	49	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	**72**	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	relevance-feedback	16	3.43	**35**	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
	q	15	3.21	7	2.81	2	1.00	1	0.67	25	12.69
	sum	467	100	249	100	200	100	149	100	198	100

Results

RQ2

		dbpedia			robust04			antique			gov2			clueweb09		
		$\#q*$	%	avg	$\#q*$	%	avg	$\#q*$	%	avg	$\#q*$	%	avg	$\#q*$	%	avg
bm25.map	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

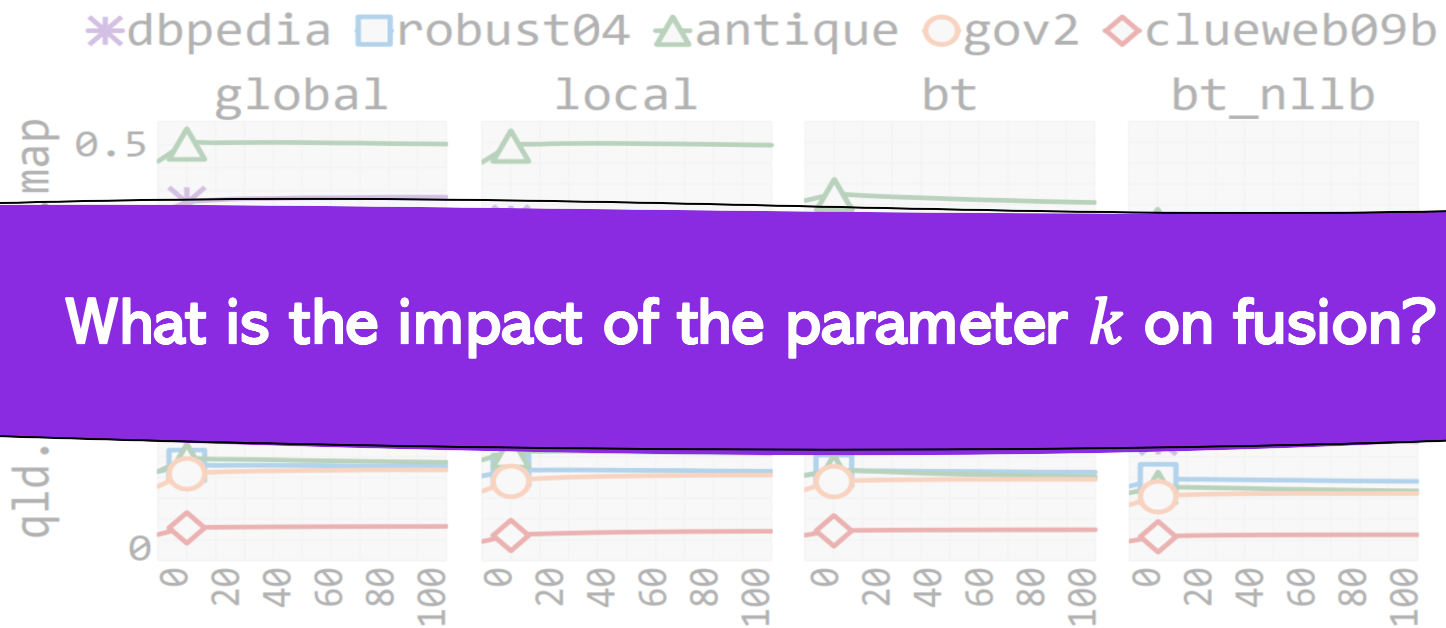
Is the effectiveness of rrf-fusion consistent across diverse datasets?

Results

RQ2

		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	**96**	20.56	0.289	**62**	24.90	0.223	37	18.50	0.404	**71**	47.65	0.231	**62**	31.31	0.088
	rrf.global	88	18.84	0.241	38	15.26	0.211	24	12.00	0.350	14	9.40	0.167	39	19.70	0.057
	rrf.local	87	18.63	0.210	46	18.47	0.183	**107**	53.50	0.239	36	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

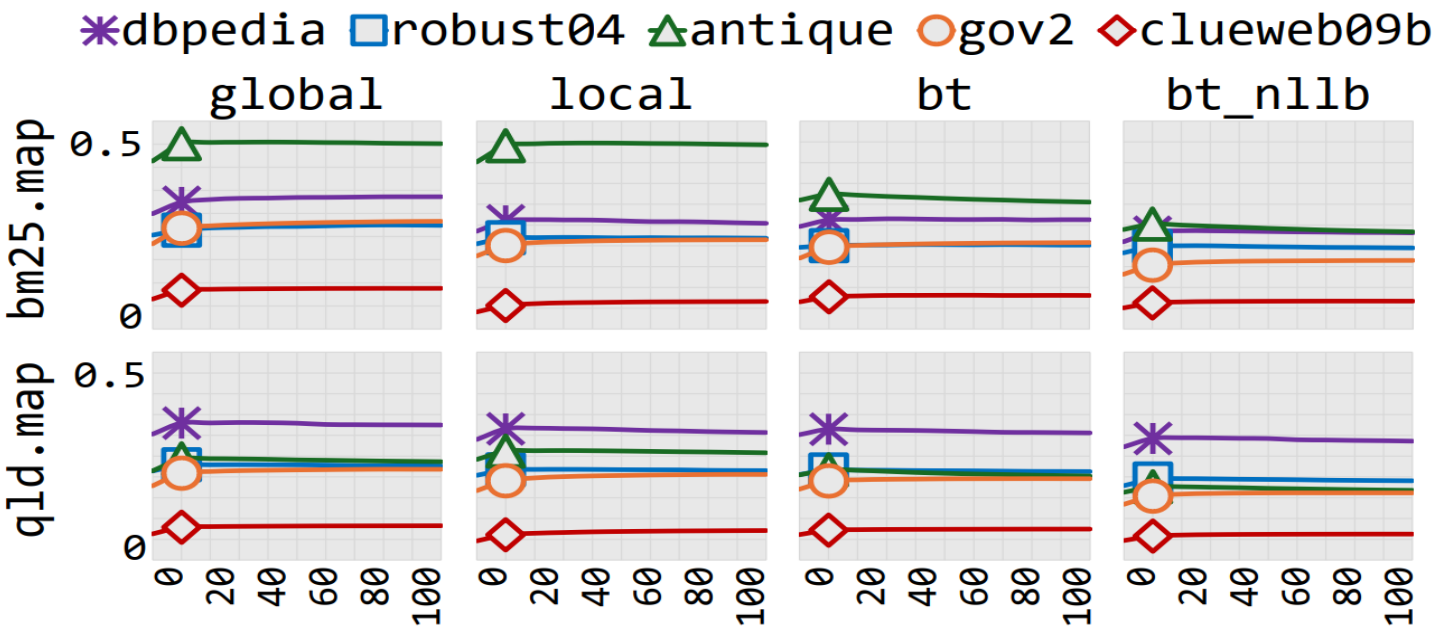
Results
RQ3



What is the impact of the parameter k on fusion?

Results

RQ3

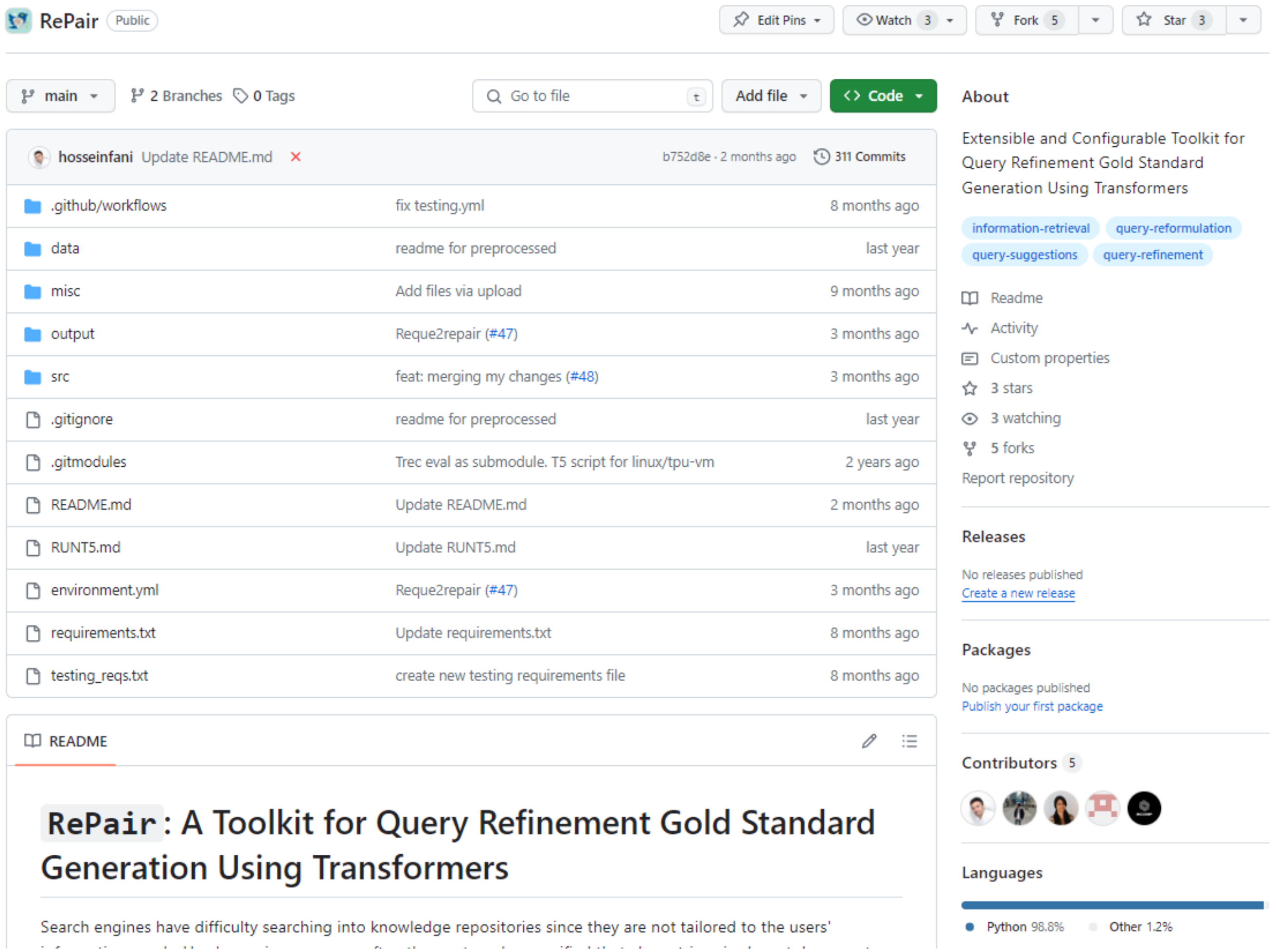


dbpedia robust04 antique gov2 clueweb09b
global
local
bt
bt_nllb
bm25.map
qld.map
0.5
0
0
20
40
60
80
100

Conclusion

RePair

We open-sourced our pipeline to support the reproducibility of our research.
https://github.com/fani-lab/RePair



Rag | Rag-fusion | Backtranslation | Methodology | Experiments | Results | Conclusion

Thank you!

Delaram Rajaei

M.Sc.
University of Windsor
rajaeid@uwindsor.ca

Zahra Taheri

Ph.D. Candidate
University of Windsor
taherik@uwindsor.ca

Hossein Fani

Assistant Professor
University of Windsor
hfani@uwindsor.ca