Zeta Alpha

A New Generation of Neural Search and Knowledge Discovery Tools

DG CNECT workshop on large language models - June 14th 2022

Jakub Zavrel

Zeta Alpha

Large Language Models will augment human cognition for making decisions.

- Human level language understanding is becoming available at scale
- Expert and executive decisions are based on knowledge
- Our capacity to discover and digest information is inherently limited

→ *Better decisions through cognitive augmentation*

Knowledge discovery

Digesting, connecting, modeling

Zeta Alpha

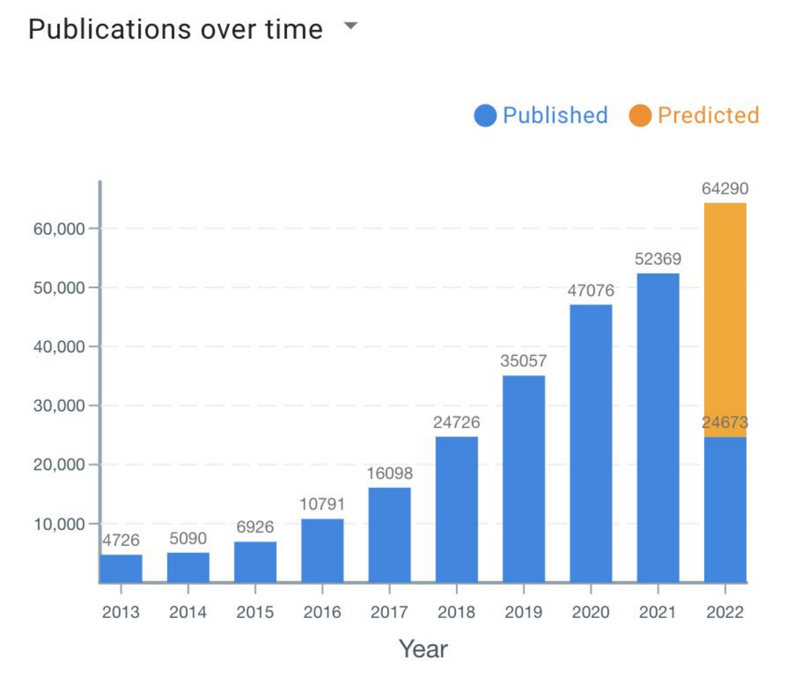
Experts in AI and Data Science

Problem:

- explosion of pre-print literature
- unknown unknowns
- connected research areas

Solution:

- personalized research assistants
- neural semantic search / recommend
- combine discovery and organization



Zeta Alpha discovery platform

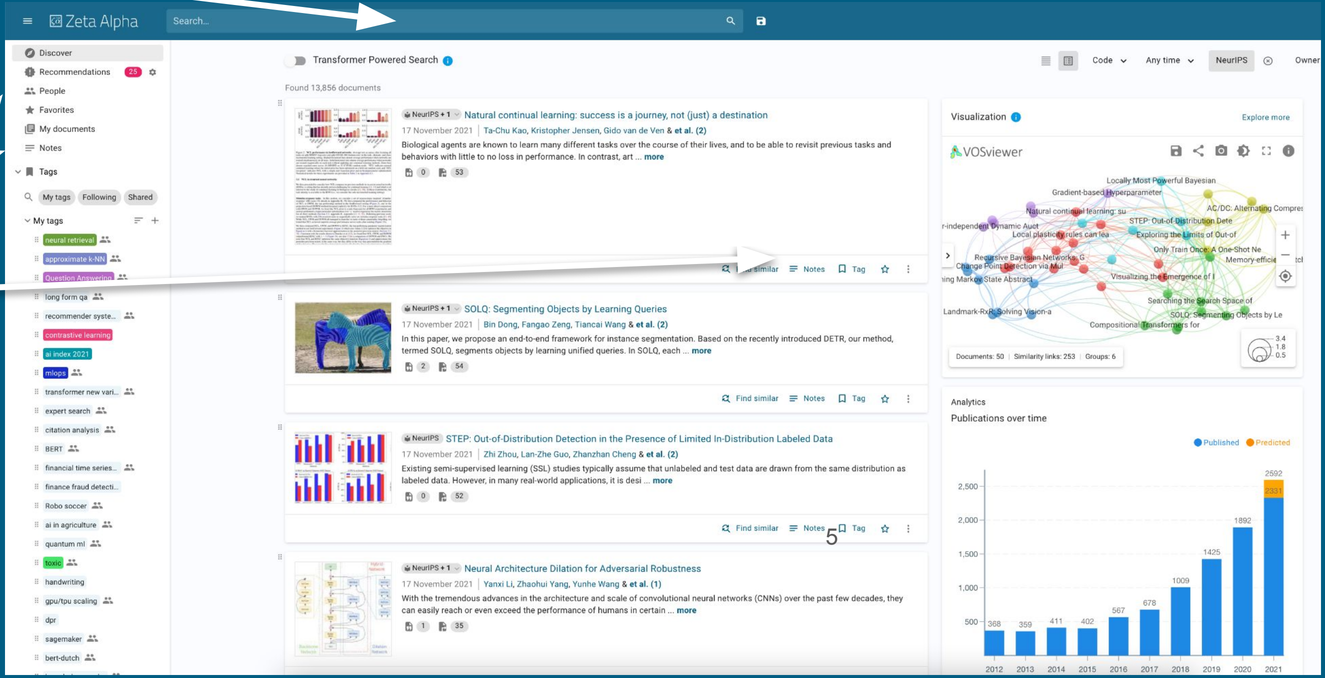
1. Discover content and people using neural search, find similar, visualization, trending on social media, and code popularity.

2. Organize your projects in personalized collections using tags.

3. Read documents, annotate and take notes.

4. Receive timely need to know recommendations tailored to your interests.

5. Share and re-use knowledge and connect within your team.



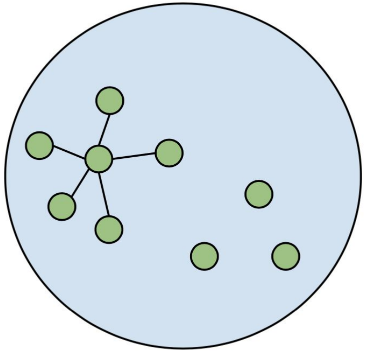
AI focused technical content from arXiv, conferences, companies, blogs, news, github code, twitter

+ import private data

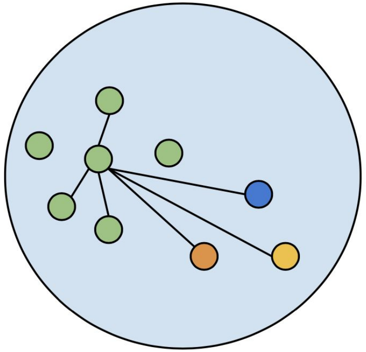
Zeta Alpha

Why Neural Search?

- Semantic understanding of data as opposed to surface keywords: bridge *the* ***lexical gap***
- Context and relationships crucial in interpreting meaning: ***handles complex and relational queries***
- Unstructured data accessible without classification and taxonomies, even ***multi-lingual*** and ***cross-lingual***
- ***Multi-modal*** capabilities: potential to combine text, audio, images and video

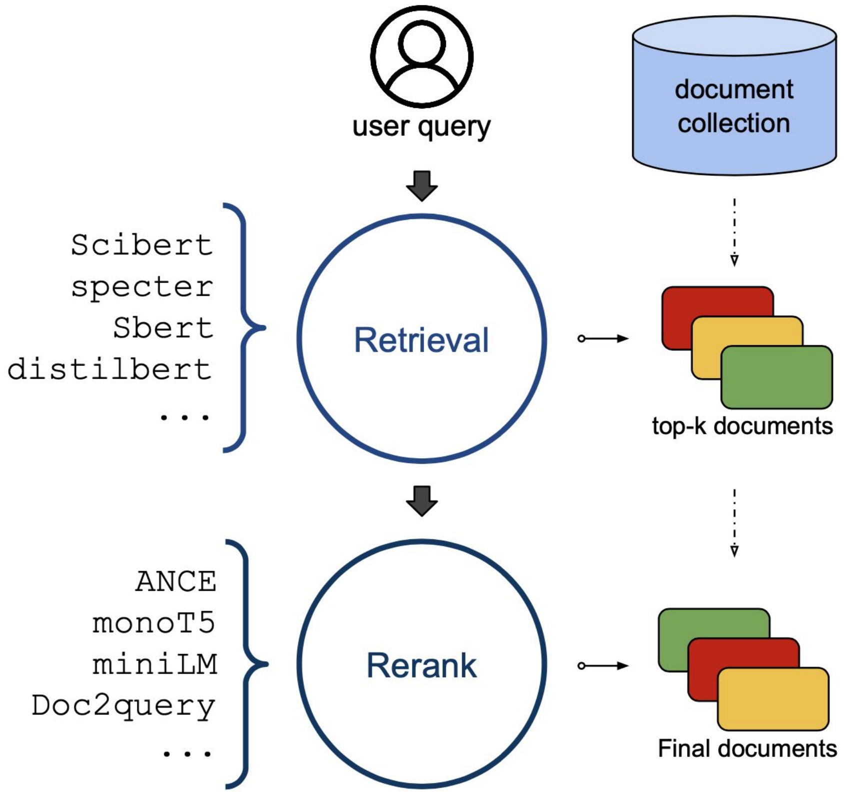


searching



expanding the horizon of exploration

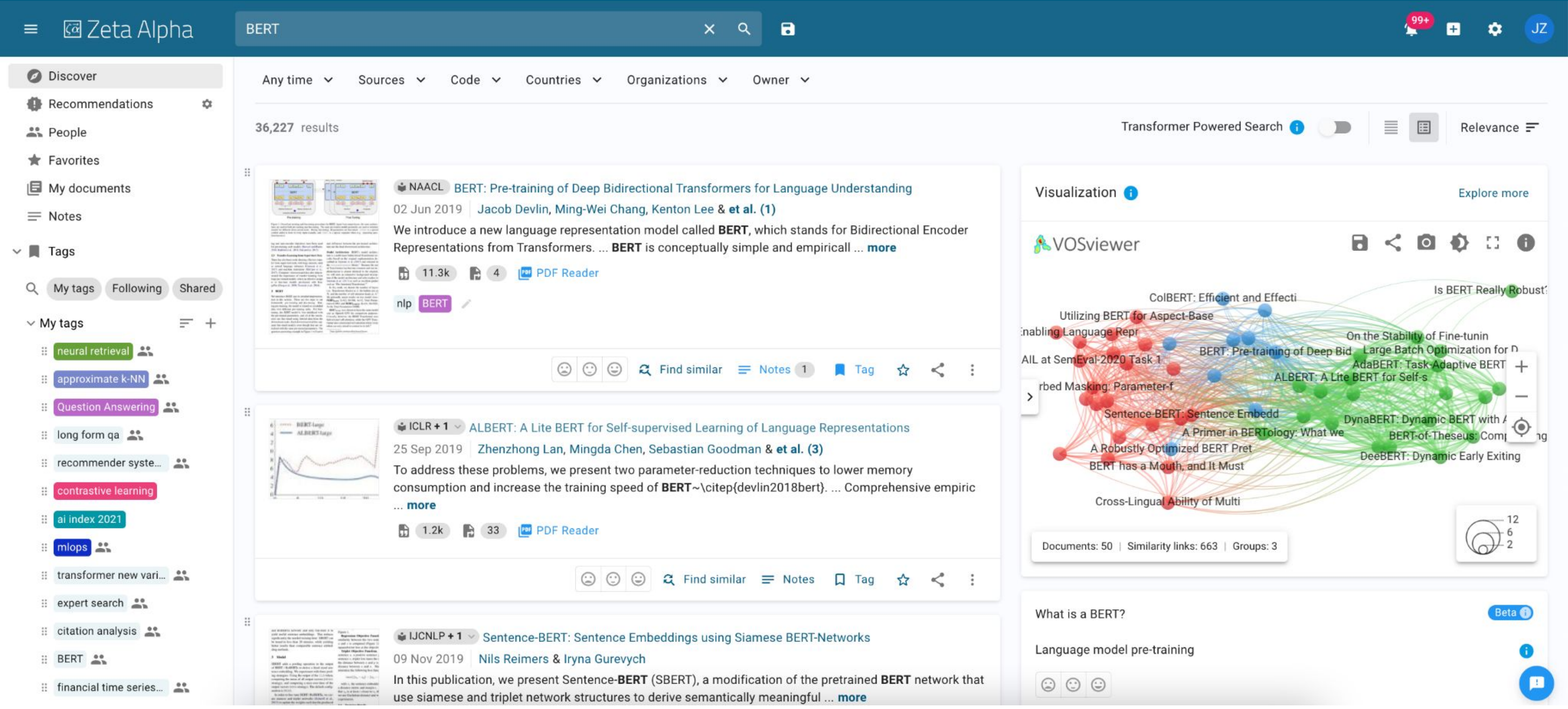
Neural Search pipeline



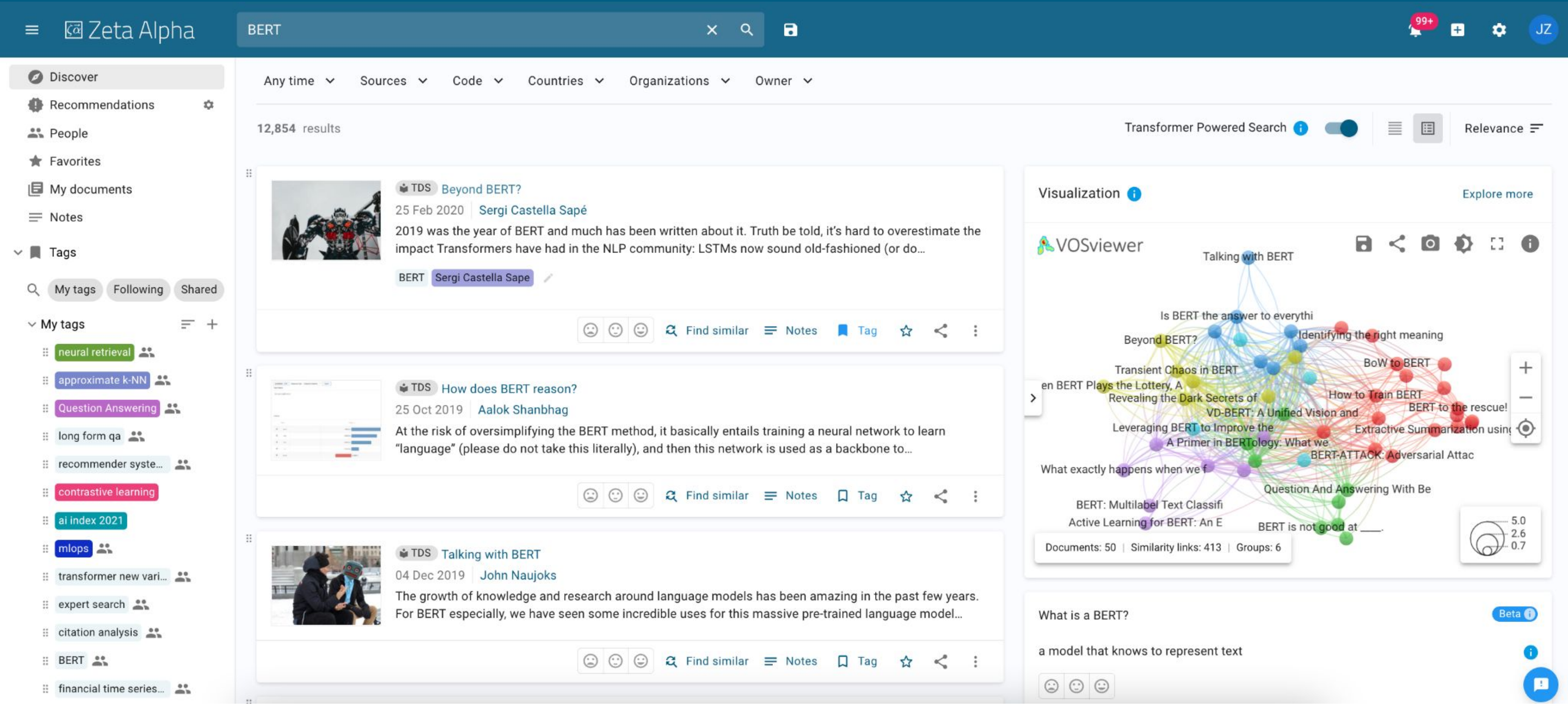
Bi-encoder + ANN search

Cross-encoder

A well tuned keyword search is hard to beat…



Neural Search, it's different…

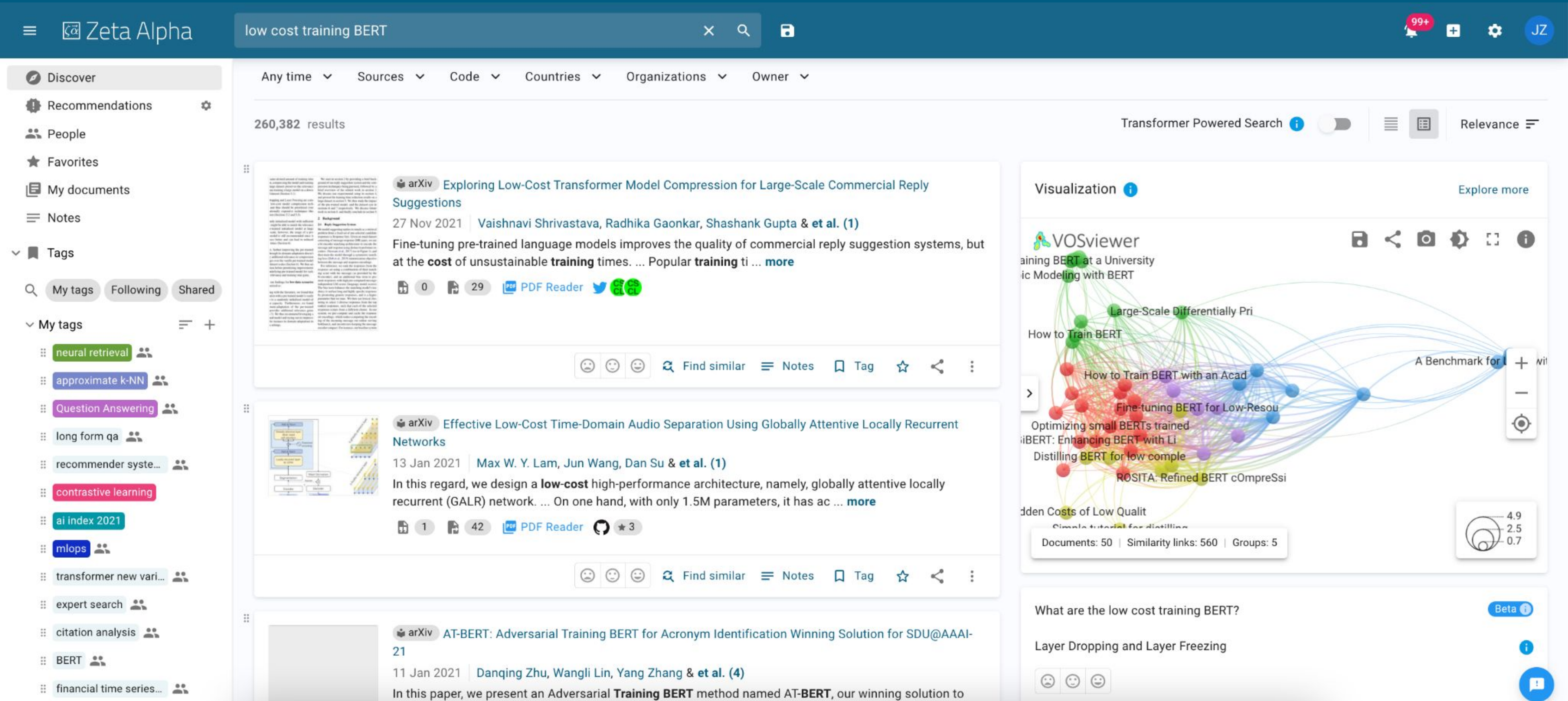


AI and Data Science domain - evaluation

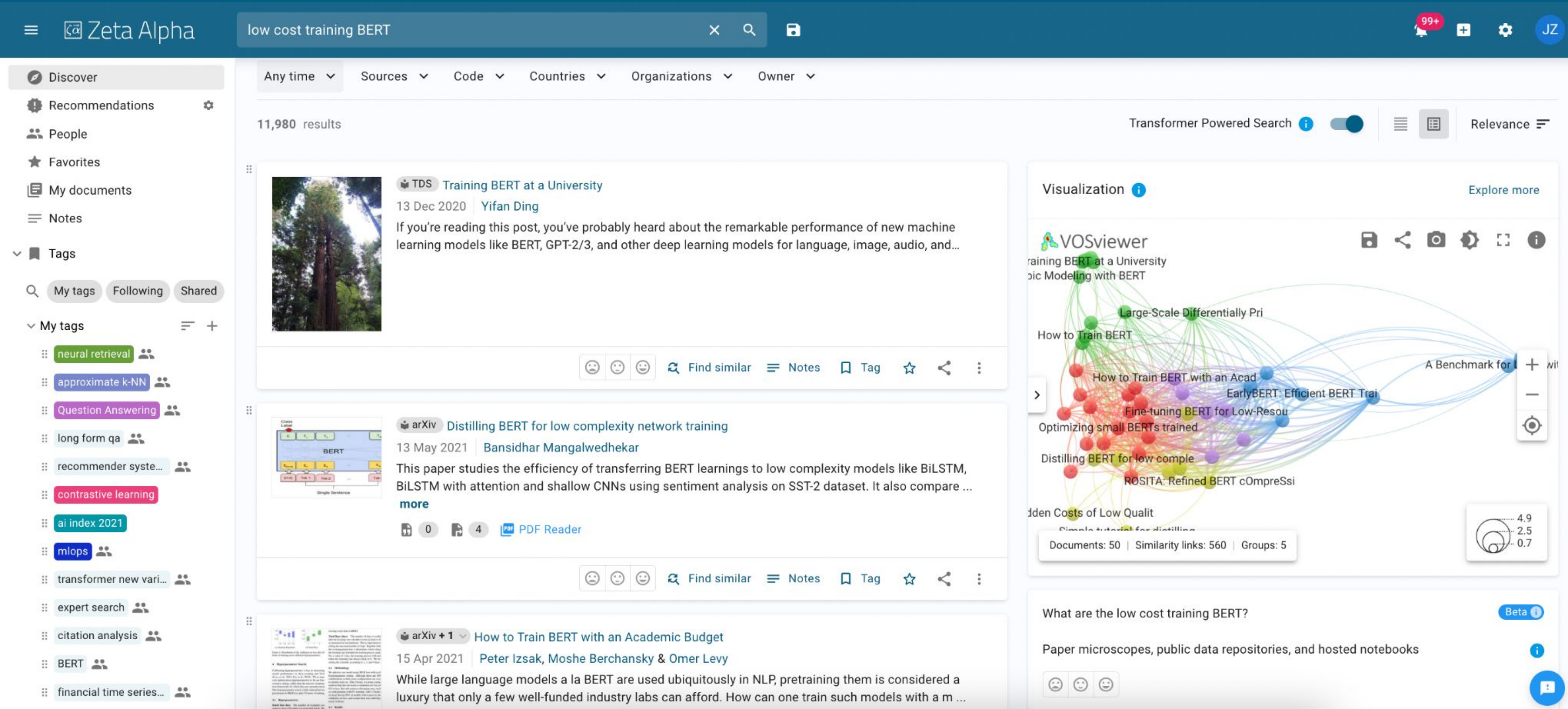
We benchmark expert queries in AI domain
(*short phrases, knowledge graph questions, quora questions, freq. user queries, and paper titles*)

Model	P@10	R@10	F1@10	MRR@10
ZA keyword search	0.71	0.17	0.27	0.91
ZA neural search	0.84	0.21	0.34	0.94

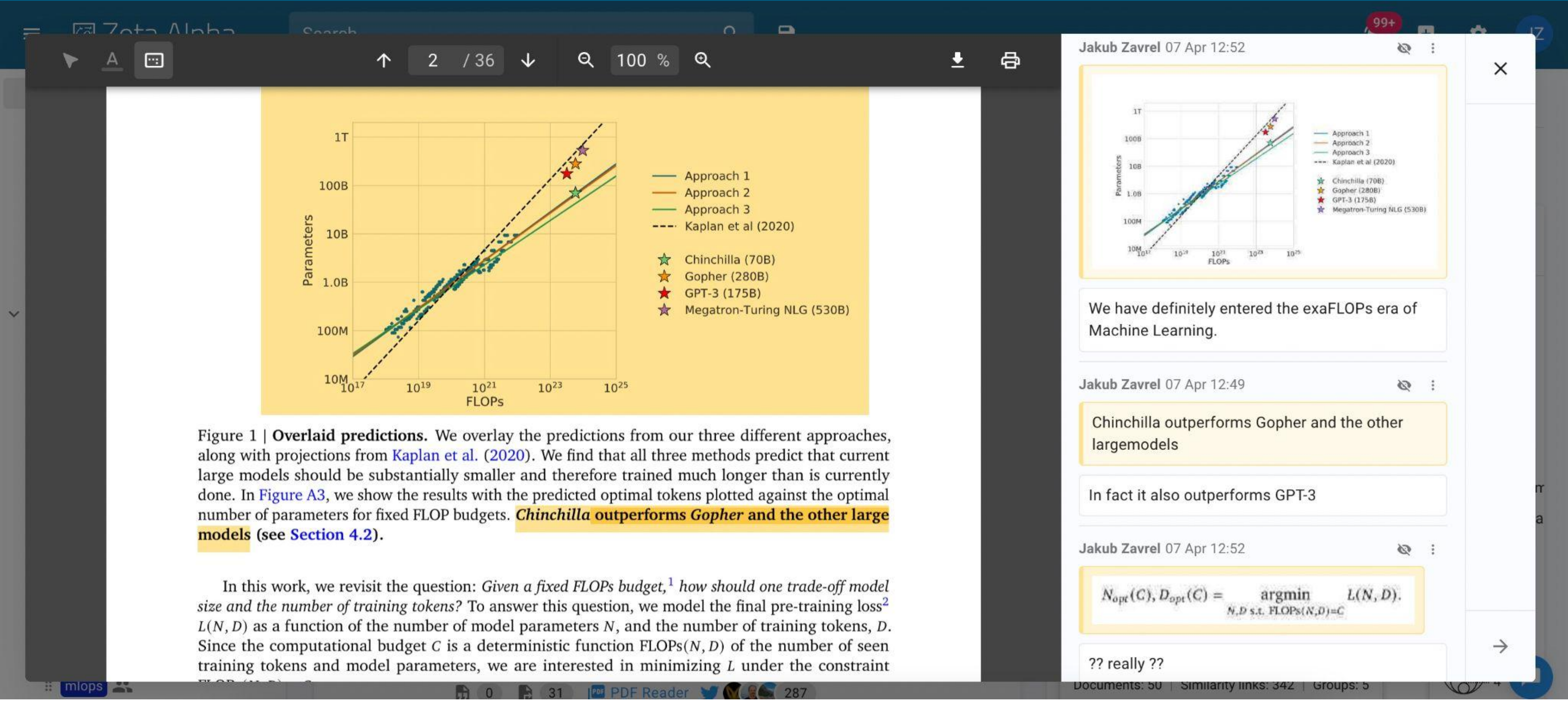
Keywords struggle on complex explorative queries



Neural Search "gets it": synonyms + relations

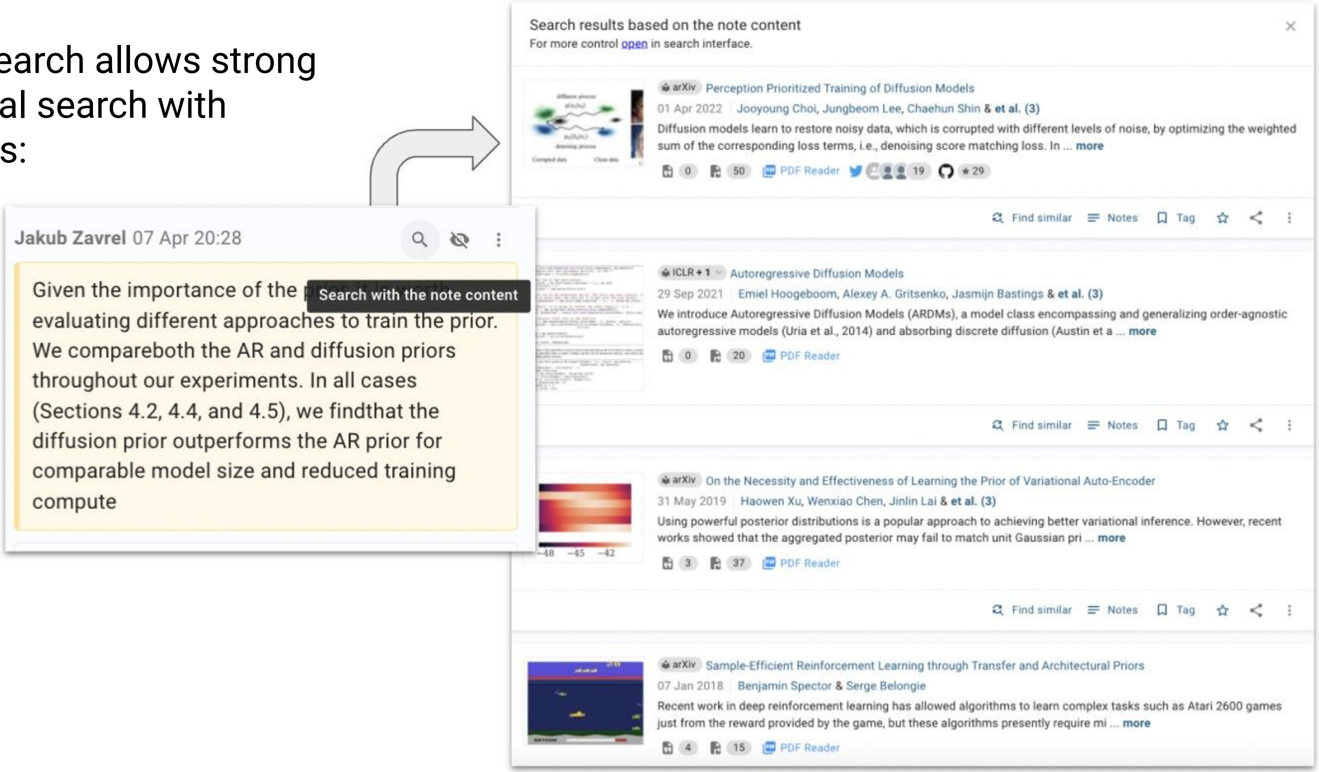


Contextual knowledge discovery

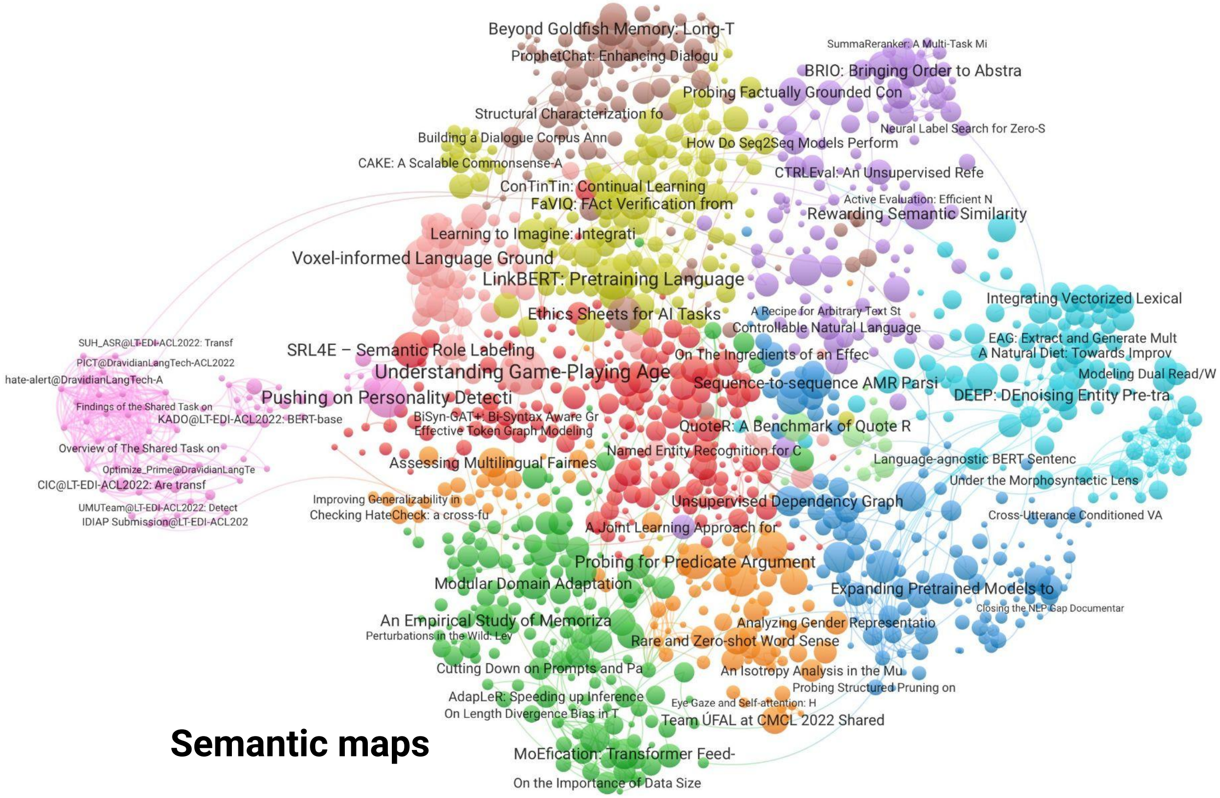


Contextual knowledge discovery

Neural Search allows strong contextual search with text notes:

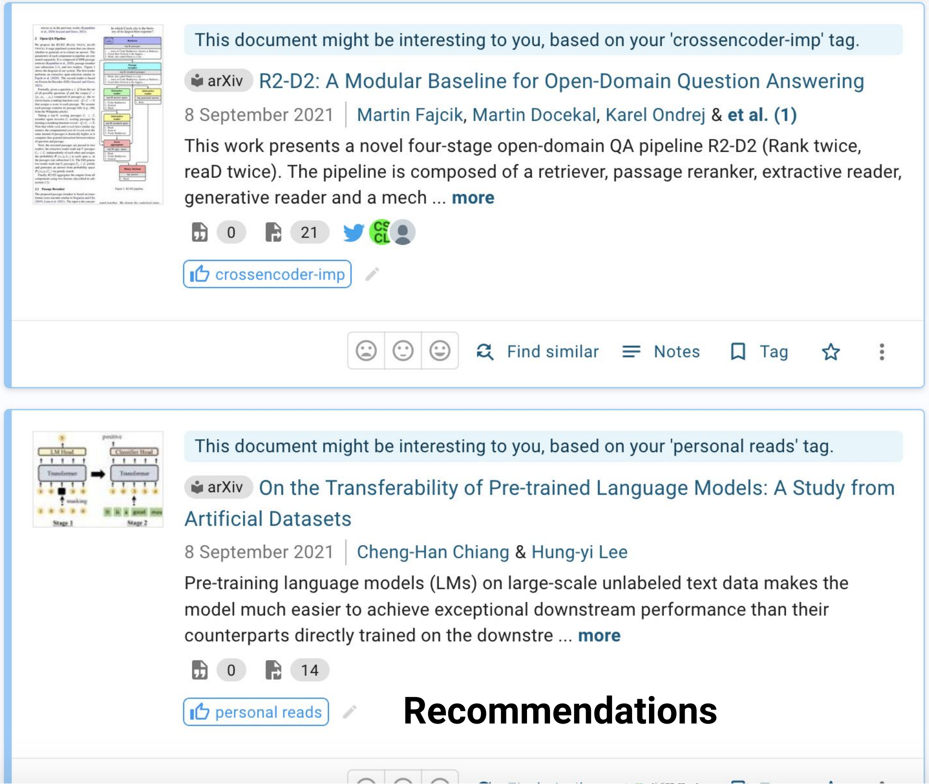


Many other uses of Neural Search

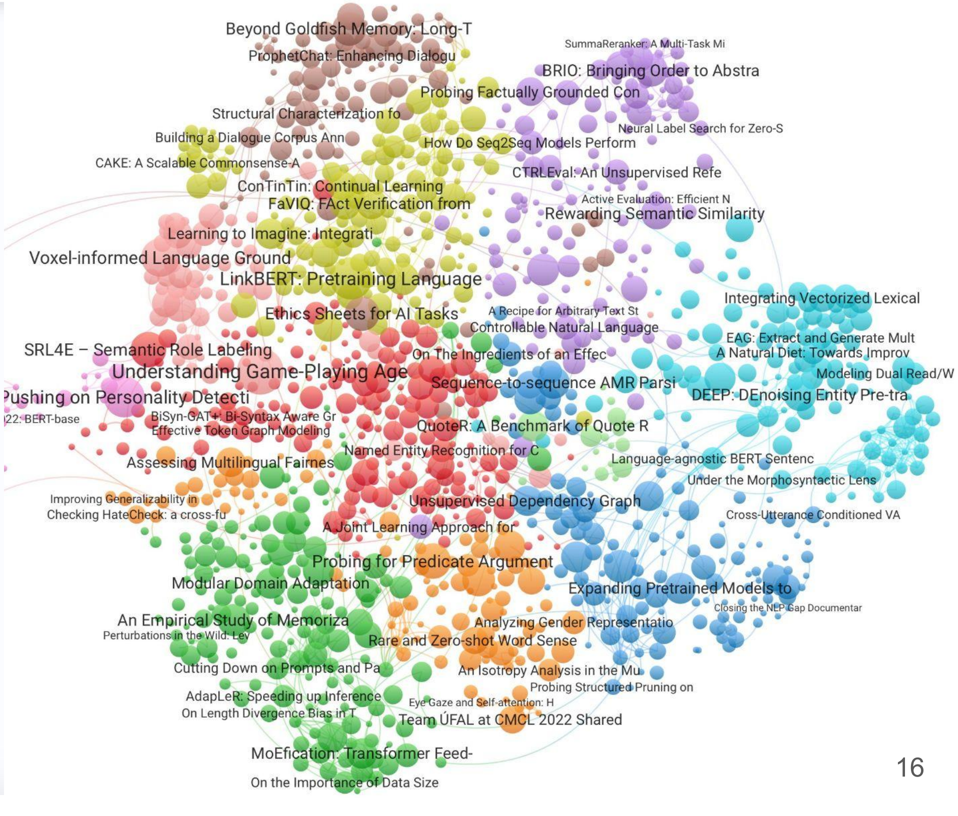


Semantic maps

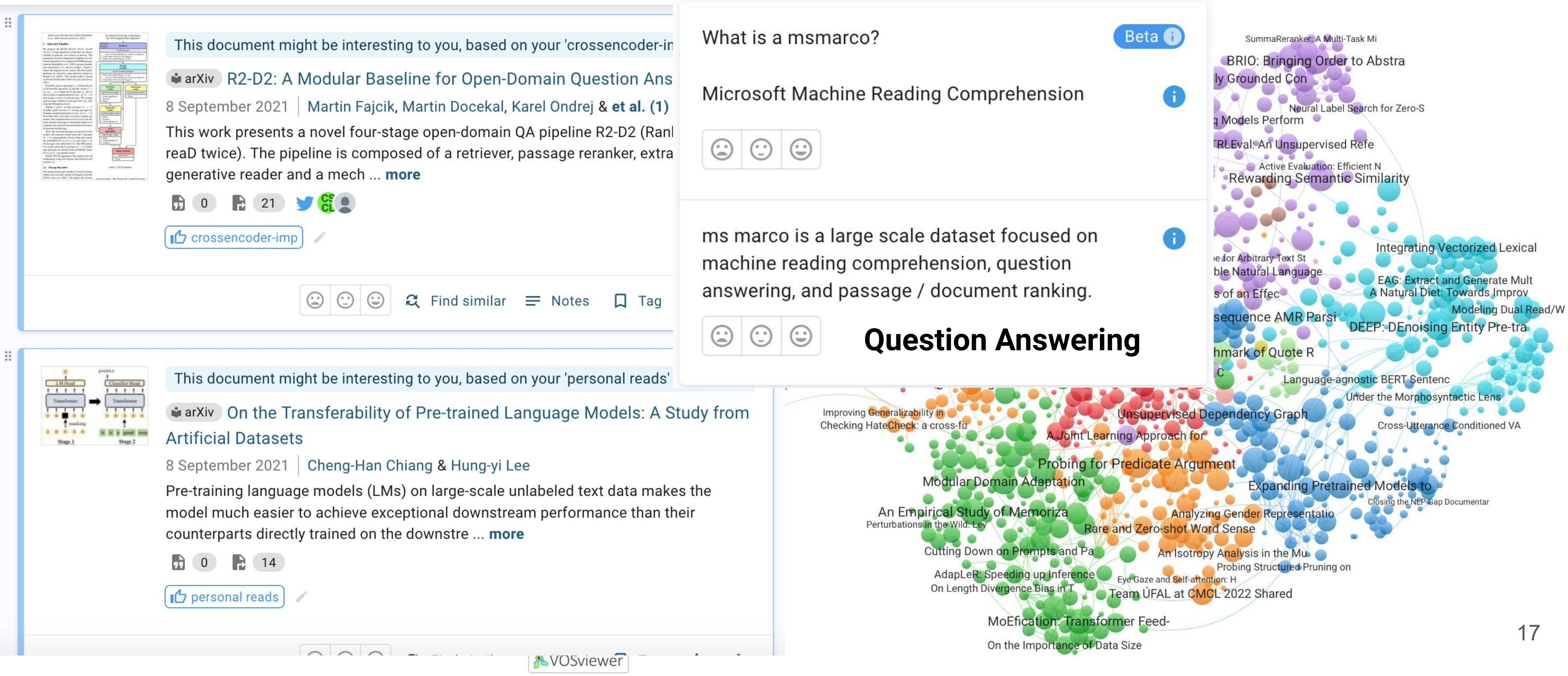
Many other uses of Neural Search



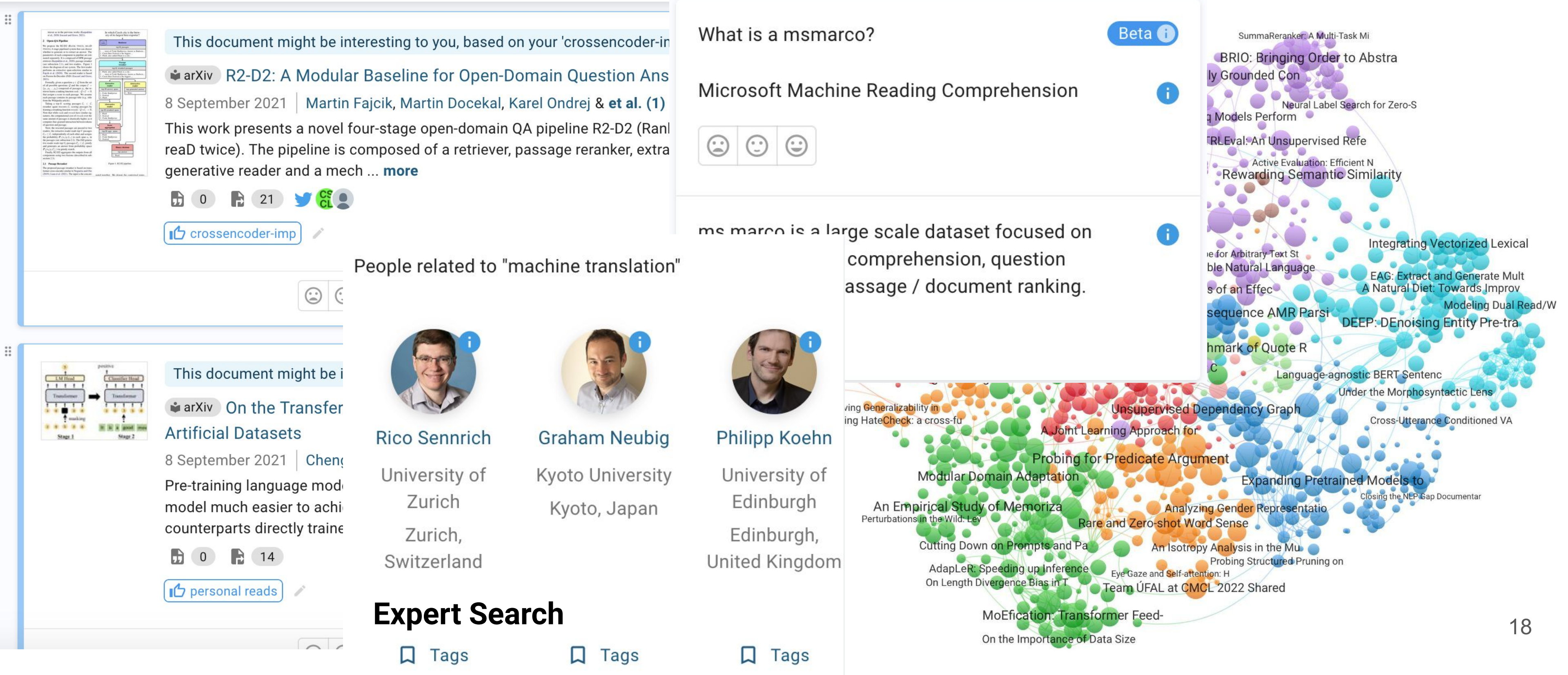
Recommendations



Many other uses of Neural Search



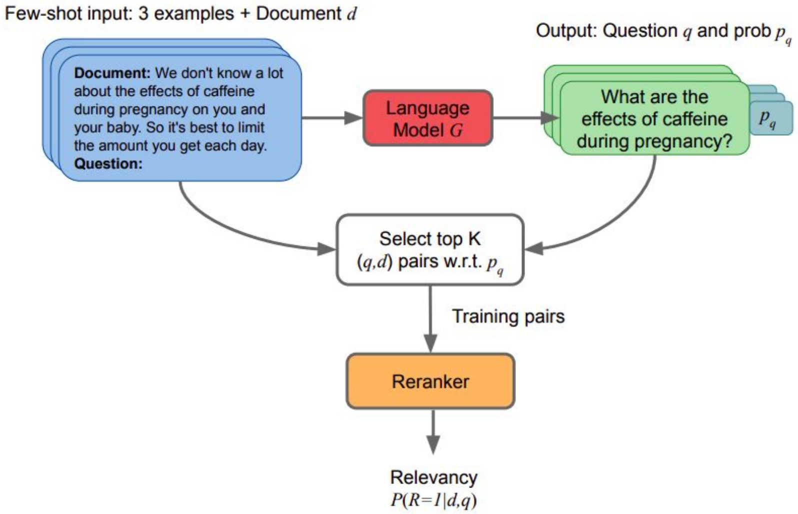
Many other uses of Neural Search

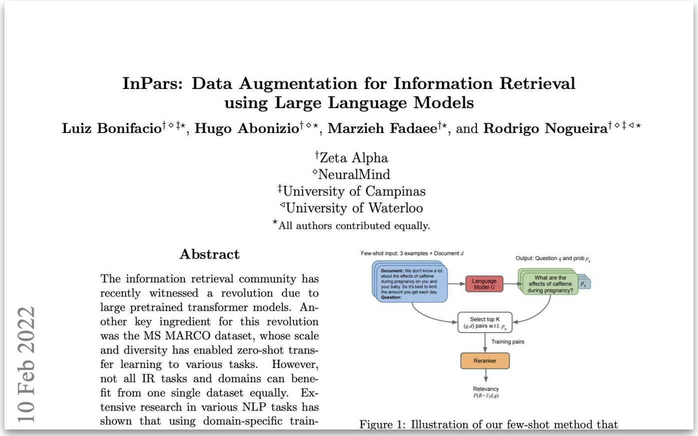


Neural Search
R&D at Zeta Alpha

1. How to adapt Neural Search to new domains without supervised training data?

InPars: Data Augmentation for Unsupervised IR





InPars: Data Augmentation for Information Retrieval using Large Language Models

Luiz Bonifacio[†⋄‡⋆], Hugo Abonizio[†⋄⋆], Marzieh Fadaee[†⋆], and Rodrigo Nogueira[†⋄‡◁⋆]

[†]Zeta Alpha
[⋄]NeuralMind
[‡]University of Campinas
[◁]University of Waterloo
[⋆]All authors contributed equally.

10 Feb 2022

Abstract

The information retrieval community has recently witnessed a revolution due to large pretrained transformer models. Another key ingredient for this revolution was the MS MARCO dataset, whose scale and diversity has enabled zero-shot transfer learning to various tasks. However, not all IR tasks and domains can benefit from one single dataset equally. Extensive research in various NLP tasks has shown that using domain-specific train-

Figure 1: Illustration of our few-shot method that

Ranking models are **finetuned on a synthetic dataset** built by augmenting **documents with queries** using **generative LLMs like GPT-3.** Our recipe for **unsupervised domain adaptation.**

InPars: Data Augmentation for IR using LLM's

		MARCO	TREC-DL 2020		Robust04		NQ	TRECC
		MRR@10	MAP	nDCG@10	MAP	nDCG@20	nDCG@10	nDCG@10
	Unsupervised							
(1)	BM25	0.1874	0.2876	0.4876	0.2531	0.4240	0.3290	0.6880
(2)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.2580	0.2740
(3)	cpt-text (Neelakantan et al., 2022)	0.2270	-	-	-	-	-	0.4270
	OpenAI Search reranking 100 docs from BM25							
(4)	Ada (300M)	\$	0.3141	0.5161	0.2691	0.4847	0.4092	0.6757
(5)	Curie (6B)	\$	0.3296	0.5422	0.2785	0.5053	0.4171	0.7251
(6)	Davinci (175B)	\$	0.3163	0.5366	0.2790	0.5103	\$	0.6918
	InPars (ours)							
(7)	monoT5-220M	0.2585	0.3599	0.5764	0.2490	0.4268	0.3354	0.6666
(8)	monoT5-3B	**0.2967**	**0.4334**	**0.6612**	**0.3180**	**0.5181**	**0.5133**	**0.7835**
	Supervised [▷ MARCO]							
(9)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.4980	0.5960
(10)	cpt-text (Neelakantan et al., 2022)	-	-	-	-	-	-	0.6490
(11)	ColBERT-v2 (Santhanam et al., 2021)	0.3970	-	-	-	-	0.5620	0.7380
(12)	GPL (Wang et al., 2021)	-	-	-	-	-	-	0.7400
(13)	miniLM reranker	†0.3901	-	-	-	-	‡0.5330	‡0.7570
(14)	monoT5-220M (Nogueira et al., 2020)	0.3810	0.4909	0.7141	0.3279	0.5298	0.5674	0.7775
(15)	monoT5-3B (Nogueira et al., 2020)	0.3980	**0.5281**	**0.7508**	0.3876	0.6091	**0.6334**	0.7948
	InPars (ours) [▷ MARCO ▷ unsup in-domain]							
(16)	monoT5-3B	0.3894	0.5087	0.7439	**0.3967**	**0.6227**	0.6297	**0.8471**

With **very good results on the BEIR benchmark…**

InPars: out of domain data augmentation

Example 1:
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee.
Relevant Query: Is a little caffeine ok during pregnancy?

Example 2:
Document: Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty, while others list the fruits as being bitter and inedible.
Relevant Query: What fruit is native to Australia?

Example 3:
Document: The Canadian Armed Forces. 1 The first large-scale Canadian peacekeeping mission started in Egypt on November 24, 1956. 2 There are approximately 65,000 Regular Force and 25,000 reservist members in the Canadian military. 3 In Canada, August 9 is designated as National Peacekeepers' Day.
Relevant Query: How large is the Canadian military?

Example 4:
Document: {document_text}
Relevant Query:

Example 1:
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee.
Good Question: How much caffeine is ok for a pregnant woman to have?
Bad Question: Is a little caffeine ok during pregnancy?

Example 2:
Document: Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty, while others list the fruits as being bitter and inedible.
Good Question: What is Passiflora herbertiana (a rare passion fruit) and how does it taste like?
Bad Question: What fruit is native to Australia?

Example 3:
Document: The Canadian Armed Forces. 1 The first large-scale Canadian peacekeeping mission started in Egypt on November 24, 1956. 2 There are approximately 65,000 Regular Force and 25,000 reservist members in the Canadian military. 3 In Canada, August 9 is designated as National Peacekeepers' Day.
Good Question: Information on the Canadian Armed Forces size and history.
Bad Question: How large is the Canadian military?

Example 4:
Document: {document_text}
Good Question:

Prompting GPT-3 by analogy

Significant improvement over OpenAI

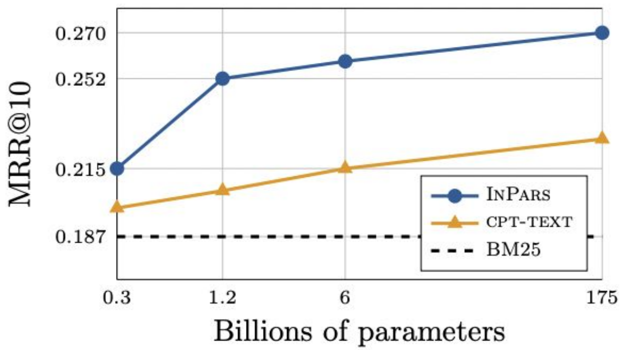


Figure 3: MRR@10 on the MS MARCO development set achieved by InPars using monoT5-220M reranker trained on synthetic questions generated by GPT-3 models of different sizes. Figures for cpt-text are from (Neelakantan et al., 2022). Note the log scale for the x-axis.

How Distillation and Size Affect Zero-Shot Retrieval

		Reranking top 1000 docs from BM25				Dense Models		
	BM25	**MiniLM**[1]	**monoT5**			**ColBERT-v2**[1]	**GTR**	**SGPT**[2]
Parameters	-	22M	60M	220M	3B	110M	4.8B	5.8B
MS MARCO	0.1870	0.3901	0.3566	0.3810	0.3980	0.3970	0.3880	-
TREC-COVID	0.5947	0.7188	0.6928	0.7775	0.7948	0.7380	0.5010	**0.8730**
NFCorpus	0.3218	0.3501	0.3180	0.3570	**0.3837**	0.3380	0.3420	0.3630
BioASQ	0.5224	0.5335	0.4880	0.5240	**0.5740**	-	0.3240	0.4130
Natural Questions	0.3055	0.5525	0.4733	0.5674	**0.6334**	0.5620	0.5680	0.5240
HotpotQA	0.6330	0.7324	0.5996	0.6950	**0.7589**	0.6670	0.5990	0.5930
FEVER	0.6513	0.8180	0.7191	0.8018	**0.8495**	0.7850	0.7400	0.7830
Climate-FEVER	0.1651	0.2555	0.2116	0.2451	0.2802	0.1760	0.2670	**0.3050**
DBPedia	0.3180	0.4652	0.3437	0.4195	**0.4777**	0.4460	0.4080	0.3990
TREC-NEWS	0.3952	0.4464	0.3848	0.4475	0.4727	-	0.3460	**0.4810**
Robust04	0.4485	0.4801	0.4222	0.5016	**0.5403**	-	0.5060	0.5140
ArguAna	0.2998	0.2941	0.0825	0.1321	0.2876	0.4630	**0.5400**	0.5140
Touché-2020	**0.4422**	0.2812	0.2643	0.2773	0.2995	0.2630	0.2560	0.2540
CQADupStack	0.2788	0.3611	0.3474	0.3808	**0.4155**	-	0.3990	0.3810
Quora	0.7886	0.8037	0.8259	0.8230	0.8407	-	**0.8920**	0.8460
SCIDOCS	0.1490	0.1629	0.1436	0.1649	**0.1970**	0.1540	0.1610	**0.1970**
SciFact	0.6789	0.6812	0.6963	0.7356	**0.7773**	0.6930	0.6620	0.7470
FiQA-2018	0.2361	0.3599	0.3377	0.4136	**0.5137**	0.3560	0.4670	0.3720
Signal-1M (RT)	**0.3304**	0.2964	0.2711	0.2771	0.3140	-	0.2730	0.2670
Average	0.4200	0.4774	0.4234	0.4745	**0.5228**	-	0.4580	0.4903
Improvement over BM25	-	0.0574	0.0034	0.0545	**0.1028**	-	0.0384	0.0703

Table 1: Results on BEIR. All results except MS MARCO are zero-shot. MS MARCO results are not included in the calculation of the average metrics. [1]Distilled models. [2]SGPT results are not completely zero-shot as the prompt was chosen based on the effectiveness in 6 datasets of the BEIR benchmark.

No Parameter Left Behind: How Distillation and Model Size Affect Zero-Shot Retrieval

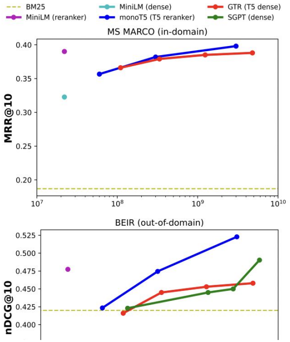
Guilherme Moraes Rosa,[1,2,3] Luiz Bonifacio,[1,2] Vitor Jeronymo,[1,2] Hugo Abonizio,[1,2] Marzieh Fadaee,[3] Roberto Lotufo,[1,2] and Rodrigo Nogueira[1,2,3]
[1]NeuralMind, Brazil
[2]UNICAMP, Brazil
[3]Zeta Alpha, Netherlands

ABSTRACT

Recent work has shown that small distilled language models are strong competitors to models that are orders of magnitude larger and slower in a wide range of information retrieval tasks. This has made distilled and dense models, due to latency constraints, the go-to choice for deployment in real-world retrieval applications. In this work, we question this practice by showing that the number of parameters and early query-document interaction play a significant role in the generalization ability of retrieval models. Our experiments show that increasing model size results in marginal gains on in-domain test sets, but much larger gains in new domains never seen during fine-tuning. Furthermore, we show that rerankers largely outperform dense ones of similar size in several tasks. Our largest reranker reaches the state of the art in 12 of the 18 datasets of the Benchmark-IR (BEIR) and surpasses the previous state of the art by 3 average points. Finally, we confirm that in-domain effectiveness is not a good indicator of zero-shot effectiveness. Code is available at https://github.com/guilhermemr04/scaling-zero-shot-retrieval.git

KEYWORDS

Distillation, Ranking, Dense retrieval, Information Retrieval, Zero-shot Learning



1 INTRODUCTION

In a more recent paper, our team has captured SOTA on the BEIR benchmark using a related approach.

2. How to adapt Neural Search to new languages without supervised training data?

mMarco: A Multilingual Version of the MS MARCO Passage Ranking Dataset

- **The MS MARCO dataset is essential for training deep learning models for Neural Search.**
- **However, MS MARCO was so far only available in English.**
- **We have now built mMARCO, a multilingual version of MS MARCO for 13 languages using machine translation.**

mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset

Luiz Henrique Bonifacio Univ. of Campinas, NeuralMind
Vitor Jeronymo Univ. of Campinas, NeuralMind
Hugo Queiroz Abonizio NeuralMind
Israel Campiotti NeuralMind
Marzieh Fadaee Zeta Alpha
Roberto Lotufo Univ. of Campinas, NeuralMind
Rodrigo Nogueira Univ. of Campinas, Univ. of Waterloo, NeuralMind

Abstract

The MS MARCO ranking dataset has been widely used for training deep learning models for IR tasks, achieving considerable effectiveness on diverse zero-shot scenarios. However, this type of resource is scarce in languages other than English. In this work, we present mMARCO, a multilingual version of the MS MARCO passage ranking dataset comprising 13 languages that was created using machine translation. We evaluated mMARCO by fine-tuning monolingual and multilingual re-ranking models, as well as a dense multilingual model on this dataset. Experimental results demonstrate that multilingual models fine-tuned on our translated

whereas for re-ranking approaches, an initial retrieval system (e.g., using a bag-of-words (BOW) or dense method) provides a list of candidates which are typically re-ranked using a cross-encoder model (Nogueira et al., 2020, 2019; Qu et al., 2021; Zhang et al., 2021b; Ma et al., 2021). Usually, the models used in both approaches are fine-tuned on a labeled dataset containing queries and examples of relevant documents.

For many languages, the available training and evaluation datasets are biased towards traditional techniques (Thakur et al., 2021), such as bag-of-words, as they are often used to build these resources (Buckley et al., 2007; Yilmaz et al., 2020). As a consequence, neural models are at a disadvan-

mMarco: A Multilingual Version of the MS MARCO Passage Ranking Dataset

	Language	R@1k BM25	R@1k mColB.	MRR@10 BM25	MRR@10 mT5	MRR@10 mMiniLM
(1)	English (Orig.)	0.857	0.953	0.184	0.366	0.366
(2)	Spanish	0.770	0.897	0.158	0.314	0.309
(3)	French	0.769	0.891	0.155	0.302	0.296
(4)	Italian	0.753	0.888	0.153	0.303	0.291
(5)	Portuguese	0.744	0.887	0.152	0.302	0.289
(6)	Indonesian	0.767	0.854	0.149	0.298	0.293
(7)	German	0.674	0.867	0.136	0.289	0.278
(8)	Russian	0.685	0.836	0.124	0.263	0.251
(9)	Chinese	0.678	0.837	0.116	0.249	0.249
Zero-shot (models were fine-tuned on the 9 languages above)						
(10)	Japanese	0.714	0.806	0.141	0.267	0.263
(11)	Dutch	0.694	0.862	0.140	0.292	0.276
(12)	Vietnamese	0.714	0.719	0.136	0.256	0.247
(13)	Hindi	0.711	0.785	0.134	0.266	0.262
(14)	Arabic	0.638	0.749	0.111	0.235	0.219

- Neural Search trained on mMARCO translated data consistently outperforms BM25 (classical keyword search).

- Fine-tuning on mMarco even allows multi-lingual LLM's to be used for neural search on unseen languages.

mMarco: A Multilingual Version of the MS MARCO Passage Ranking Dataset

	Translation	*es*	*fr*	*pt*	*it*	*id*	*de*	*ru*	*zh*	*ar*	*hi*	**avg**
mT5												
(5)	Helsinki*	0.297	0.279	0.285	0.248	0.244	0.264	0.183	0.152	0.187	0.035	0.217
(6)	Google	0.314	0.302	0.302	0.303	0.298	0.289	0.263	0.249	0.235	0.266	0.281

Table 3: Comparison of Helsinki translation models (open source) vs Google Translate (commercial). The reported metric is MRR@10 on the development set of mMARCO.

** Helsinki - Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In Proc. of EAMT, Lisbon, Portugal.*

mMARCO is available from https://github. com/unicamp-dl/mMARCO.git.

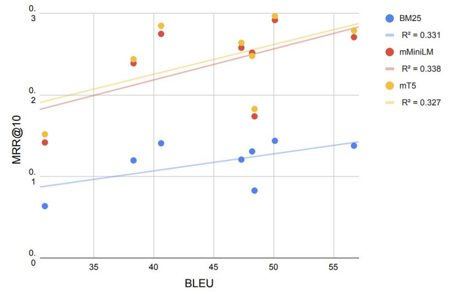


Figure 1: Translation quality measured as BLEU on Tatoeba vs retrieval quality measured as MRR@10 on mMARCO.

Summary:

1. Large language models open the doors to a new generation of knowledge discovery and cognitive augmentation tools that will lead towards cognitive augmentation of expert decision making.

2. The general availability of powerful language models for text generation and translation allows the flourishing of the European AI industry, by enabling the creation of synthetic in-domain training data in all European languages.

Zeta Alpha

Any Questions?

Publications

- **InPars: Data Augmentation for Information Retrieval using Large Language Models**
  2022 | Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee & Rodrigo Nogueira
  See paper
- **Building a Platform for Ensemble-based Personalized Research Literature Recommendations for AI and Data Science at Zeta Alpha**
  2021 | Jakub Zavrel, Artem Grotov, & Jonathan Mitnik
  See paper
- **mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset**
  2021 | Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo & Rodrigo Nogueira
  See paper
- **Pretrained Transformers for Text Ranking: BERT and Beyond**
  2021 | Jimmy Lin, Rodrigo Nogueira, & Andrew Yates
  Access book
- **A New Neural Search and Insights Platform for Navigating and Organizing AI Research**
  2020 | Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp, Jakub Zavrel
  See paper
- **Effective Distributed Representations for Academic Expert Search**
  2020 | Mark Berger, Jakub Zavrel, & Paul Groth
  See paper

Get more information, sign up to use the platform:

www.zeta-alpha.com