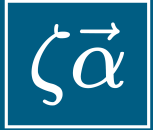


Zeta Alpha

A New Generation of Neural Search and Knowledge Discovery Tools

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

Jakub Zavrel



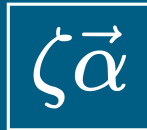
Large Language Models will augment human cognition for making decisions.

- 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



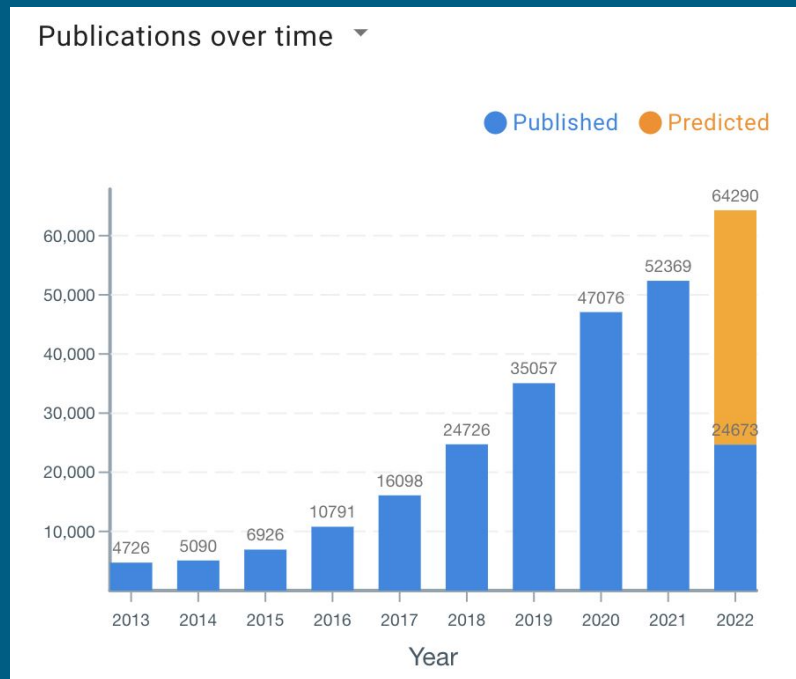
Experts in AI and Data Science

Problem:

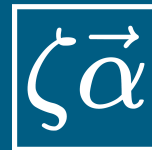
- explosion of pre-print literature
- unknown unknowns
- connected research areas

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.

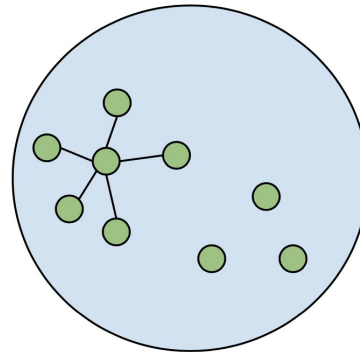
The screenshot shows the Zeta Alpha discovery platform interface. The main content area displays search results for 'Transformer Powered Search', showing 13,856 documents. Three document cards are visible, each with a thumbnail, title, date, authors, and a brief description. The left sidebar contains navigation options: Discover, Recommendations (63), People, Favorites, My documents, Notes, and Tags. The right sidebar includes a Visualization section with a network graph and an Analytics section with a bar chart showing 'Publications over time' from 2012 to 2021.

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

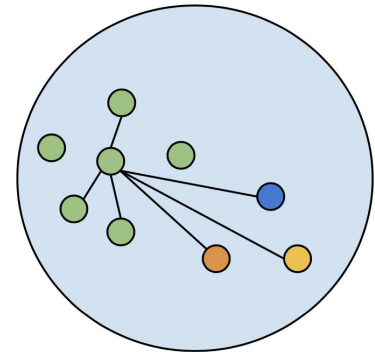
+ import private data

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

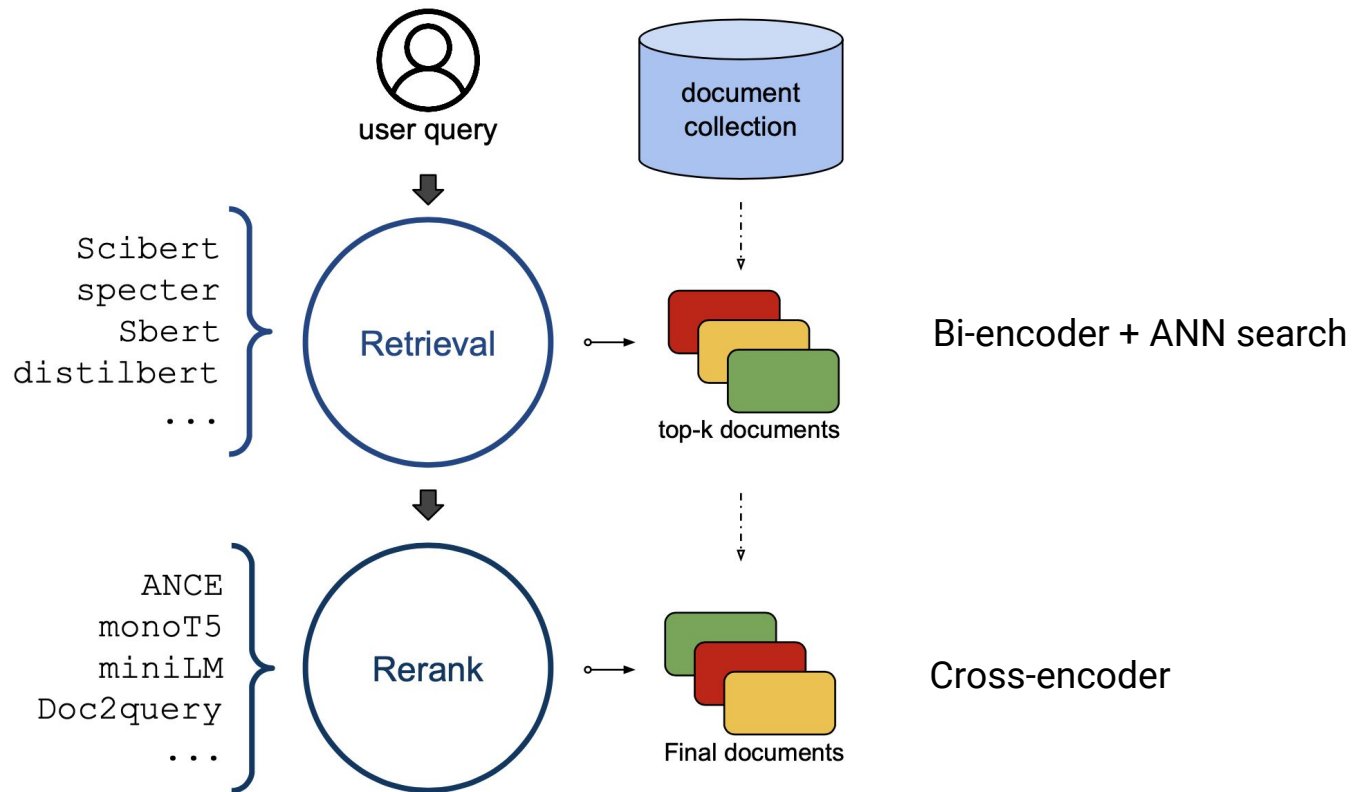


searching



expanding the
horizon of
exploration

Neural Search pipeline



A well tuned keyword search is hard to beat...



Navigation bar for Zeta Alpha search interface. Includes a search bar with the text "BERT", a magnifying glass icon, and a close button. On the right, there are icons for user profile (JZ), settings, and a notification badge showing "99+".

Left sidebar navigation menu. It includes sections for "Discover", "Recommendations", "People", "Favorites", "My documents", "Notes", "Tags", "My tags", and "My tags" (repeated). Under "My tags", there are several tags: "neural retrieval", "approximate k-NN", "Question Answering", "long form qa", "recommender syste...", "contrastive learning", "ai index 2021", "mlops", "transformer new vari...", "expert search", "citation analysis", "BERT", and "financial time series...".

Main search results area. At the top, it shows "36,227 results" and filter options for "Any time", "Sources", "Code", "Countries", "Organizations", and "Owner". The results are displayed in a list format. The first result is titled "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Jacob Devlin, Ming-Wei Chang, and Kenton Lee, dated 02 Jun 2019. It has 11.3k views and 4 PDFs. The second result is "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations" by Zhenzhong Lan, Mingda Chen, Sebastian Goodman, et al., dated 25 Sep 2019. It has 1.2k views and 33 PDFs. The third result is "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" by Nils Reimers & Iryna Gurevych, dated 09 Nov 2019. It has 1.2k views and 33 PDFs. Each result includes a thumbnail image, a brief description, and a "more" link.

Right sidebar area. At the top, it says "Transformer Powered Search" with a toggle switch and a "Relevance" filter. Below this is a "Visualization" section with a "VOSviewer" logo and a network graph. The graph shows various research topics related to BERT, such as "ColBERT: Efficient and Effective", "Utilizing BERT for Aspect-Based", "BERT: Pre-training of Deep Bidirectional", "ALBERT: A Lite BERT for Self-supervised Learning", "Sentence-BERT: Sentence Embeddings", "A Primer in BERTology: What We", "A Robustly Optimized BERT Pre-training", "BERT has a Mouth, and It Must", "Cross-Lingual Ability of Multi", "DynaBERT: Dynamic BERT with Adaptive", "BERT-of-Theseus: Compressing", "DeeBERT: Dynamic Early Exiting", "Large Batch Optimization for", "AdaBERT: Task-Adaptive BERT", and "On the Stability of Fine-tuning". Below the graph, it shows "Documents: 50 | Similarity links: 663 | Groups: 3". At the bottom, there is a "What is a BERT?" section with a "Beta" badge and a "Language model pre-training" section.



Neural Search, it's different...

☰ Zeta Alpha

BERT

✕ 🔍 📄

99+



JZ

Discover

Recommendations

People

Favorites

My documents

Notes

Tags

My tags Following Shared

My tags

neural retrieval

approximate k-NN

Question Answering

long form qa

recommender syste...

contrastive learning

ai index 2021

mlops

transformer new vari...

expert search

citation analysis

BERT

financial time series...

Any time Sources Code Countries Organizations Owner

12,854 results



TDS Beyond BERT?

25 Feb 2020 Sergi Castella Sapé

2019 was the year of BERT and much has been written about it. Truth be told, it's hard to overestimate the impact Transformers have had in the NLP community: LSTMs now sound old-fashioned (or do...

BERT Sergi Castella Sape

Find similar Notes Tag ☆



TDS How does BERT reason?

25 Oct 2019 Aalok Shanbhag

At the risk of oversimplifying the BERT method, it basically entails training a neural network to learn "language" (please do not take this literally), and then this network is used as a backbone to...

Find similar Notes Tag ☆



TDS Talking with BERT

04 Dec 2019 John Naujoks

The growth of knowledge and research around language models has been amazing in the past few years. For BERT especially, we have seen some incredible uses for this massive pre-trained language model...

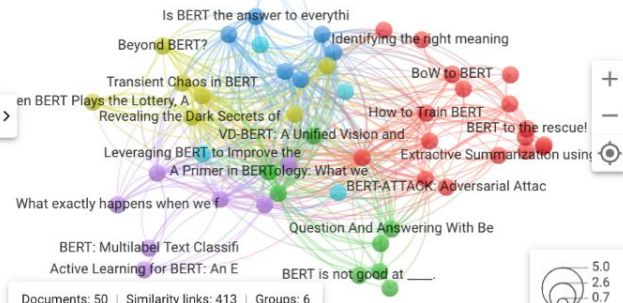
Find similar Notes Tag ☆

Transformer Powered Search Relevance

Visualization Explore more

VOSviewer

Talking with BERT



What is a BERT?

a model that knows to represent text

Find similar Notes Tag ☆

Beta

i

🗨



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

Discover low cost training BERT

260,382 results

Any time Sources Code Countries Organizations Owner

Transformer Powered Search Relevance

Discover Recommendations People Favorites My documents Notes Tags My tags Following Shared My tags

- neural retrieval
- approximate k-NN
- Question Answering
- long form qa
- recommender syste...
- contrastive learning
- ai index 2021
- mlops
- transformer new vari...
- expert search
- citation analysis
- BERT
- financial time series...

arXiv Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions

27 Nov 2021 Vaishnavi Shrivastava, Radhika Gaonkar, Shashank Gupta & et al. (1)

Fine-tuning pre-trained language models improves the quality of commercial reply suggestion systems, but at the **cost** of unsustainable **training** times. ... Popular **training** ti ... [more](#)

0 PDF Reader

Find similar Notes Tag ☆

arXiv Effective Low-Cost Time-Domain Audio Separation Using Globally Attentive Locally Recurrent Networks

13 Jan 2021 Max W. Y. Lam, Jun Wang, Dan Su & et al. (1)

In this regard, we design a **low-cost** high-performance architecture, namely, globally attentive locally recurrent (GALR) network. ... On one hand, with only 1.5M parameters, it has ac ... [more](#)

1 PDF Reader

Find similar Notes Tag ☆

arXiv AT-BERT: Adversarial Training BERT for Acronym Identification Winning Solution for SDU@AAAI-21

11 Jan 2021 Danqing Zhu, Wangli Lin, Yang Zhang & et al. (4)

In this paper, we present an Adversarial **Training BERT** method named AT-BERT, our winning solution to

Visualization

VOSviewer

Training BERT at a University
Acoustic Modeling with BERT

Large-Scale Differentially Pri
How to Train BERT
How to Train BERT with an Acad
Fine-tuning BERT for Low-Resou
Optimizing small BERTs trained
Distilling BERT for low comple
ROSITA: Refined BERT cOmpreSsi

A Benchmark for

Documents: 50 | Similarity links: 560 | Groups: 5

4.9
2.5
0.7

What are the low cost training BERT?

Layer Dropping and Layer Freezing

Beta



Neural Search “gets it”: synonyms + relations


Discover

- Recommendations
- People
- Favorites
- My documents
- Notes
- Tags
- My tags: Following, Shared
- My tags: neural retrieval, approximate k-NN, Question Answering, long form qa, recommender syste..., contrastive learning, ai index 2021, mlops, transformer new vari..., expert search, citation analysis, BERT, financial time series...

Any time Sources Code Countries Organizations Owner

11,980 results

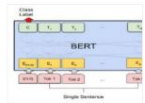
TDS Training BERT at a University
13 Dec 2020 | Yifan Ding



If you're reading this post, you've probably heard about the remarkable performance of new machine learning models like BERT, GPT-2/3, and other deep learning models for language, image, audio, and...

Find similar Notes Tag ☆ Share

arXiv Distilling BERT for low complexity network training
13 May 2021 | Bansidhar Mangalwedhekar



This paper studies the efficiency of transferring BERT learnings to low complexity models like BiLSTM, BiLSTM with attention and shallow CNNs using sentiment analysis on SST-2 dataset. It also compare ...

more

0 PDF Reader

Find similar Notes Tag ☆ Share

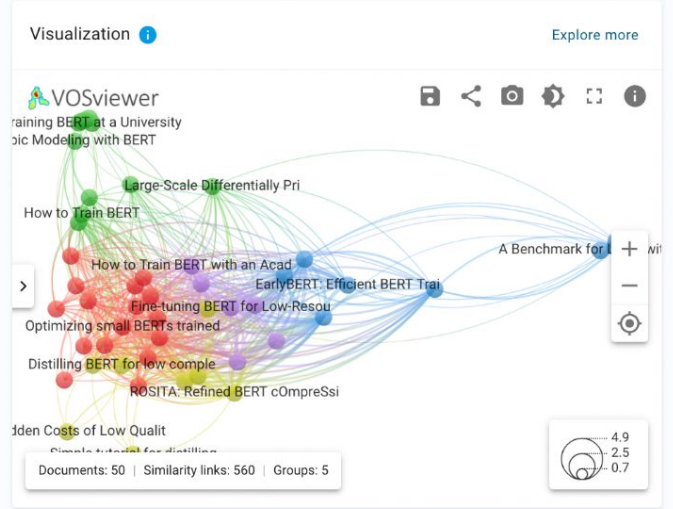
arXiv + 1 How to Train BERT with an Academic Budget
15 Apr 2021 | Peter Izsak, Moshe Berchansky & Omer Levy



While large language models like BERT are used ubiquitously in NLP, pretraining them is considered a luxury that only a few well-funded industry labs can afford. How can one train such models with a m ...

Find similar Notes Tag ☆ Share

Transformer Powered Search Relevance



What are the low cost training BERT? **Beta**

Paper microscopes, public data repositories, and hosted notebooks

Find similar Notes Tag ☆ Share

Contextual knowledge discovery

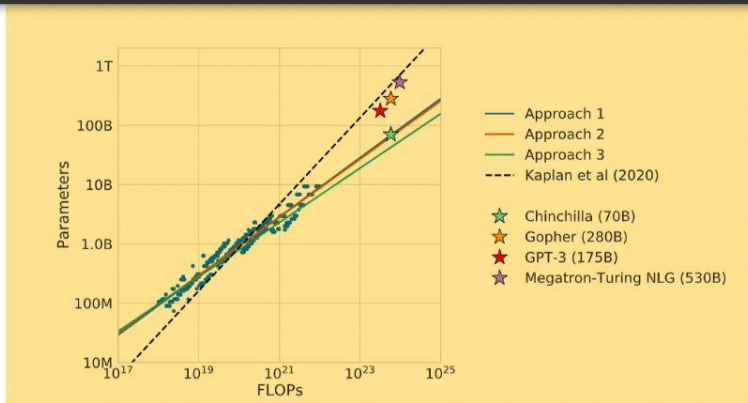
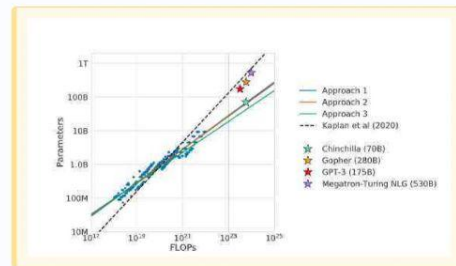


Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from [Kaplan et al. \(2020\)](#). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In [Figure A3](#), we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. **Chinchilla outperforms Gopher and the other large models** (see [Section 4.2](#)).

In this work, we revisit the question: *Given a fixed FLOPs budget,¹ how should one trade-off model size and the number of training tokens?* To answer this question, we model the final pre-training loss² $L(N, D)$ as a function of the number of model parameters N , and the number of training tokens, D . Since the computational budget C is a deterministic function $\text{FLOPs}(N, D)$ of the number of seen training tokens and model parameters, we are interested in minimizing L under the constraint

Jakub Zavrel 07 Apr 12:52



We have definitely entered the exaFLOPs era of Machine Learning.

Jakub Zavrel 07 Apr 12:49

Chinchilla outperforms Gopher and the other largemodels

In fact it also outperforms GPT-3

Jakub Zavrel 07 Apr 12:52

$$N_{\text{opt}}(C), D_{\text{opt}}(C) = \underset{N, D \text{ s.t. } \text{FLOPs}(N, D) = C}{\text{argmin}} L(N, D).$$

?? really ??



Contextual knowledge discovery

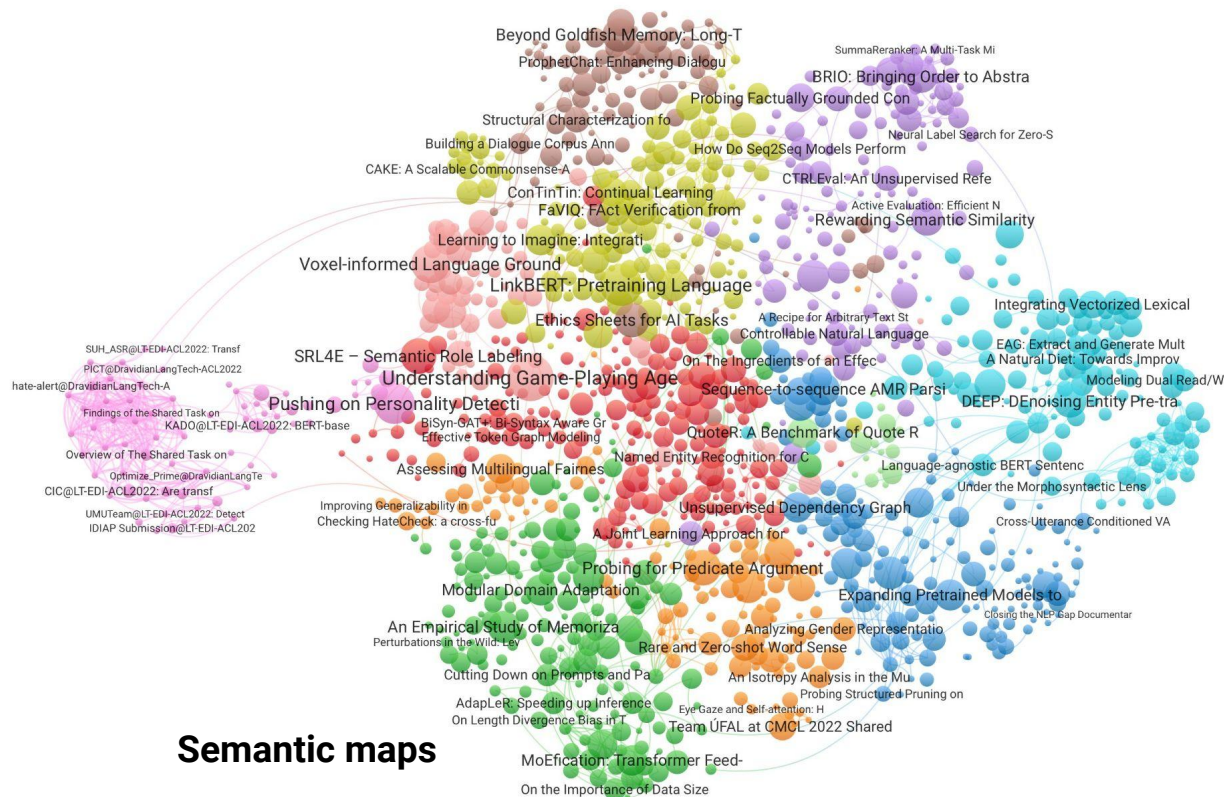
Neural Search allows strong contextual search with text notes:

The diagram illustrates the process of contextual knowledge discovery. On the left, a note from 'Jakub Zavrel' dated '07 Apr 20:28' is shown. The note contains the text: 'Given the importance of the evaluating different approaches to train the prior. We compare both the AR and diffusion priors throughout our experiments. In all cases (Sections 4.2, 4.4, and 4.5), we find that the diffusion prior outperforms the AR prior for comparable model size and reduced training compute'. A search bar within the note interface contains the text 'Search with the note content'. A large grey arrow points from this search bar to a search results window on the right. The search results window is titled 'Search results based on the note content' and lists several relevant papers:

- arXiv** Perception Prioritized Training of Diffusion Models
01 Apr 2022 Jooyoung Choi, Jungbeom Lee, Chaehun Shin & et al. (3)
Diffusion models learn to restore noisy data, which is corrupted with different levels of noise, by optimizing the weighted sum of the corresponding loss terms, i.e., denoising score matching loss. In ... more
- ICLR * 1** Autoregressive Diffusion Models
29 Sep 2021 Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings & et al. (3)
We introduce Autoregressive Diffusion Models (ARDMs), a model class encompassing and generalizing order-agnostic autoregressive models (Urie et al., 2014) and absorbing discrete diffusion (Austin et al. ... more
- arXiv** On the Necessity and Effectiveness of Learning the Prior of Variational Auto-Encoder
31 May 2019 Haowen Xu, Wenxiao Chen, Jinlin Lai & et al. (3)
Using powerful posterior distributions is a popular approach to achieving better variational inference. However, recent works showed that the aggregated posterior may fail to match unit Gaussian pri ... more
- arXiv** Sample-Efficient Reinforcement Learning through Transfer and Architectural Priors
07 Jan 2018 Benjamin Spector & Serge Belongie
Recent work in deep reinforcement learning has allowed algorithms to learn complex tasks such as Atari 2600 games just from the reward provided by the game, but these algorithms presently require ml ... more



Many other uses of Neural Search





Many other uses of Neural Search

This document might be interesting to you, based on your 'crossencoder-imp' tag.

arXiv R2-D2: A Modular Baseline for Open-Domain Question Answering
8 September 2021 | Martin Fajcik, Martin Docekal, Karel Ondrej & et al. (1)

This work presents a novel four-stage open-domain QA pipeline R2-D2 (Rank twice, read twice). The pipeline is composed of a retriever, passage reranker, extractive reader, generative reader and a mech ... [more](#)

0 21

[crossencoder-imp](#)

Find similar Notes Tag ☆

This document might be interesting to you, based on your 'personal reads' tag.

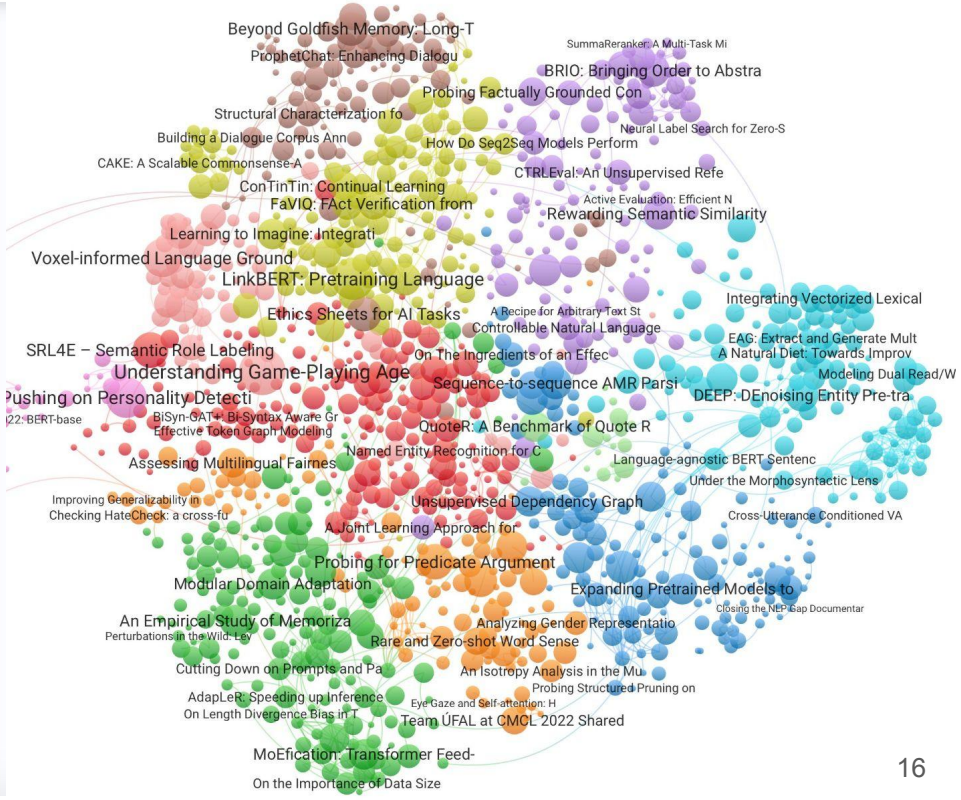
arXiv On the Transferability of Pre-trained Language Models: A Study from Artificial Datasets
8 September 2021 | Cheng-Han Chiang & Hung-yi Lee

Pre-training language models (LMs) on large-scale unlabeled text data makes the model much easier to achieve exceptional downstream performance than their counterparts directly trained on the downstre ... [more](#)

0 14

[personal reads](#)

Recommendations





Many other uses of Neural Search

This document might be interesting to you, based on your 'crossencoder-ir'

R2-D2: A Modular Baseline for Open-Domain Question Answering
8 September 2021 | Martin Fajcik, Martin Docekal, Karel Ondrej & et al. (1)

This work presents a novel four-stage open-domain QA pipeline R2-D2 (Ran, read twice). The pipeline is composed of a retriever, passage reranker, extra generative reader and a mech ... [more](#)

0 21

[crossencoder-imp](#)

Find similar Notes

This document might be interesting to you, based on your 'personal reads'

On the Transferability of Pre-trained Language Models: A Study from Artificial Datasets
8 September 2021 | Cheng-Han Chiang & Hung-yi Lee

Pre-training language models (LMs) on large-scale unlabeled text data makes the model much easier to achieve exceptional downstream performance than their counterparts directly trained on the downstre ... [more](#)

0 14

[personal reads](#)

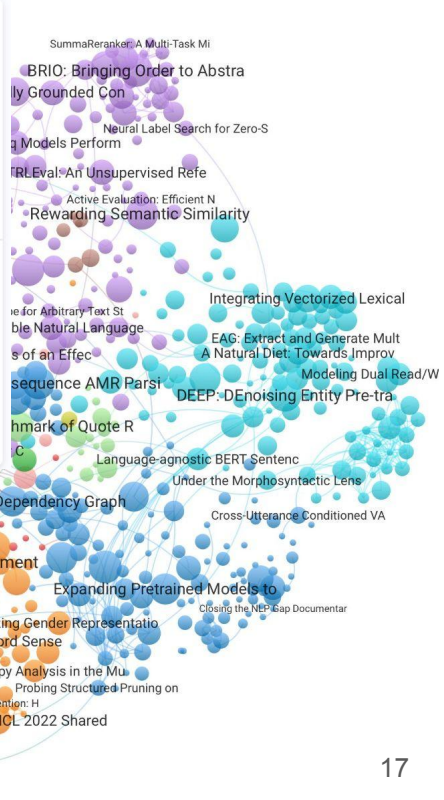
What is a msmarco?

Beta

Microsoft Machine Reading Comprehension

ms marco is a large scale dataset focused on machine reading comprehension, question answering, and passage / document ranking.

Question Answering





Many other uses of Neural Search

This document might be interesting to you, based on your 'crossencoder-ir'

R2-D2: A Modular Baseline for Open-Domain Question Answering
 8 September 2021 | Martin Fajcik, Martin Docekal, Karel Ondrej & et al. (1)
 This work presents a novel four-stage open-domain QA pipeline R2-D2 (Ran read twice). The pipeline is composed of a retriever, passage reranker, extra generative reader and a mech ... [more](#)

0 21

[crossencoder-imp](#)

This document might be interesting to you, based on your 'personal reads'

On the Transferability of Pre-trained Language Models to Artificial Datasets
 8 September 2021 | Chenyang Lu
 Pre-training language models are much easier to achieve than their counterparts directly trained on natural language. This work presents a novel four-stage open-domain QA pipeline R2-D2 (Ran read twice). The pipeline is composed of a retriever, passage reranker, extra generative reader and a mech ... [more](#)

0 14

[personal reads](#)

People related to "machine translation"



Rico Sennrich
 University of Zurich
 Zurich, Switzerland



Graham Neubig
 Kyoto University
 Kyoto, Japan



Philipp Koehn
 University of Edinburgh
 Edinburgh, United Kingdom

Expert Search

Tags

Tags

Tags

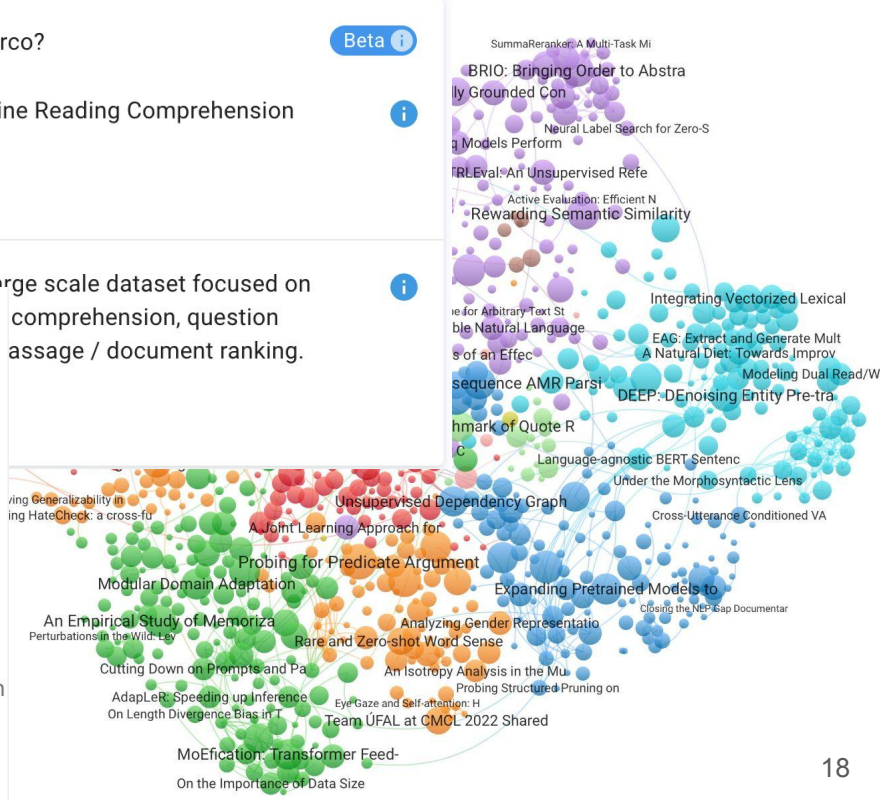
What is a msMarco?

Microsoft Machine Reading Comprehension



ms Marco is a large scale dataset focused on comprehension, question passage / document ranking.

Beta

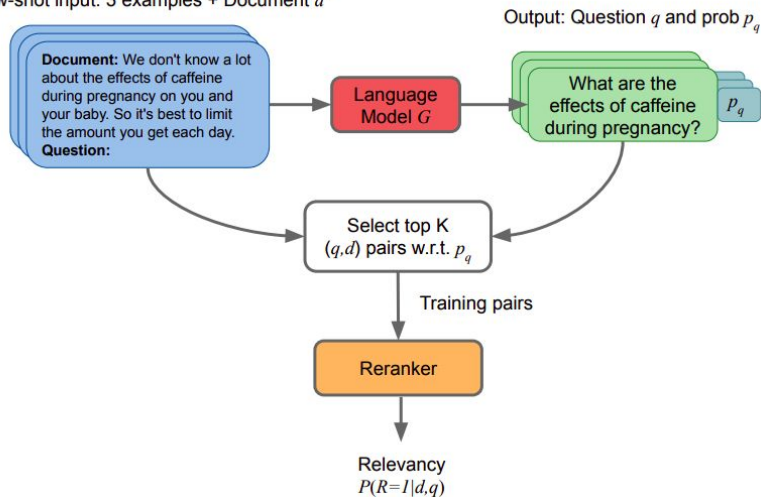


Neural Search R&D at Zeta Alpha

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

InPars: Data Augmentation for Unsupervised IR

Few-shot input: 3 examples + Document d



10 Feb 2022

InPars: Data Augmentation for Information Retrieval using Large Language Models

Luiz Bonifacio^{1,2,3}, Hugo Abonizio^{1,3*}, Marzieh Fadaee^{1*}, and Rodrigo Nogueira^{1,2,3*}

¹Zeta Alpha
²NeuralMind
³University of Campinas
⁴University of Waterloo
*All authors contributed equally.

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
<i>Unsupervised</i>							
(1)	BM25	0.1874	0.2876	0.4876	0.2531	0.4240	0.3290
(2)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.2580
(3)	cpt-text (Neelakantan et al., 2022)	0.2270	-	-	-	-	0.4270
<i>OpenAI Search reranking 100 docs from BM25</i>							
(4)	Ada (300M)	\$	0.3141	0.5161	0.2691	0.4847	0.4092
(5)	Curie (6B)	\$	0.3296	0.5422	0.2785	0.5053	0.4171
(6)	Davinci (175B)	\$	0.3163	0.5366	0.2790	0.5103	\$ 0.6918
<i>InPars (ours)</i>							
(7)	monoT5-220M	0.2585	0.3599	0.5764	0.2490	0.4268	0.3354
(8)	monoT5-3B	0.2967	0.4334	0.6612	0.3180	0.5181	0.5133
<i>Supervised</i> [\triangleright MARCO]							
(9)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.4980
(10)	cpt-text (Neelakantan et al., 2022)	-	-	-	-	-	0.6490
(11)	ColBERT-v2 (Santhanam et al., 2021)	0.3970	-	-	-	-	0.5620
(12)	GPL (Wang et al., 2021)	-	-	-	-	-	0.7400
(13)	miniLM reranker	\dagger 0.3901	-	-	-	-	\ddagger 0.5330
(14)	monoT5-220M (Nogueira et al., 2020)	0.3810	0.4909	0.7141	0.3279	0.5298	0.5674
(15)	monoT5-3B (Nogueira et al., 2020)	0.3980	0.5281	0.7508	0.3876	0.6091	0.6334
<i>InPars (ours)</i> [\triangleright MARCO \triangleright 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:

Significant improvement over OpenAI

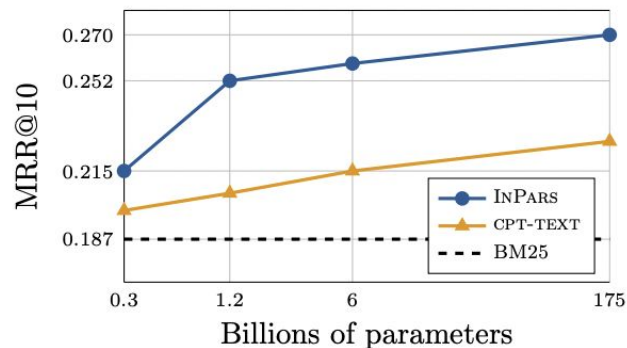


Figure 3: MRR@10 on the MS MARCO development set achieved by InPars using monoT5-220M reranker trained on synthetic questions generated by GPT-3 models of different sizes. Figures for cpt-text are from (Neelakantan et al., 2022). Note the log scale for the x-axis.

Prompting GPT-3 by analogy



How Distillation and Size Affect Zero-Shot Retrieval

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

Parameters	BM25	Reranking top 1000 docs from BM25				Dense Models		
		MiniLM ¹	monoT5		ColBERT-v2 ¹	GTR	SGPT ²	
	-	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
QQADupStack	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. ¹Distilled models. ²SGPT 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}
¹NeuralMind, Brazil
²UNICAMP, Brazil
³Zeta Alpha, Netherlands

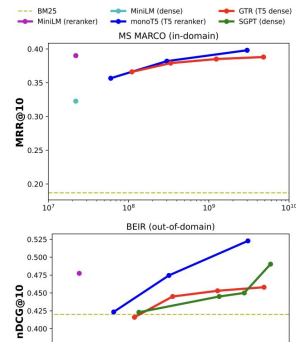
ABSTRACT

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-

397v4 [cs.CL] 10 Jan 2022

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



	Language	R@1k		MRR@10		
		BM25	mColB.	BM25	mT5	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
<i>Zero-shot (models were fine-tuned on the 9 languages above)</i>						
(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	<i>es</i>	<i>fr</i>	<i>pt</i>	<i>it</i>	<i>id</i>	<i>de</i>	<i>ru</i>	<i>zh</i>	<i>ar</i>	<i>hi</i>	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>

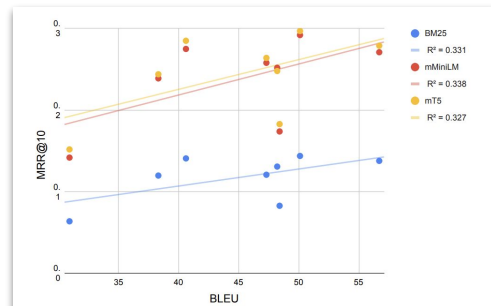


Figure 1: Translation quality measured as BLEU on Tatoeba vs retrieval quality measured as MRR@10 on mMARCO.

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.

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