MRL 2024 @ EMNLP 2024

MLT-DR: Multi-Lingual/Task Demonstration Retrieval
An Attempt towards Generalized Retriever for In-Context Learning

Retriever models will be available at:
https://github.com/google-deepmind/mlt_dr

Google

Kazuma Hashimoto, Arjun Reddy Akula, Karthik Raman, Michael Bendersky {kazumah, arjunakula, karthikraman, bemike}@google.com

TL;DR

- Investigating **generalization ability** of retrieval for in-context learning (ICL)
- Training with **81 datasets with diverse tasks, domains, and languages**
- Augmenting the training data with Google Translate **for >230 languages**

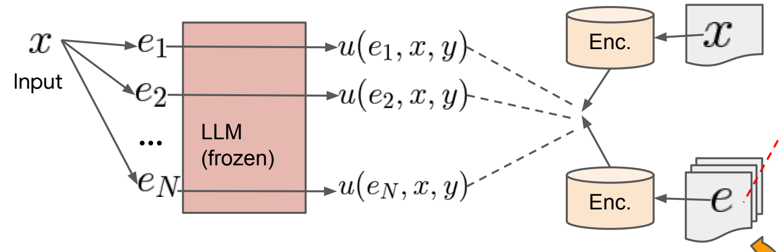
[Many tasks]

translation, NLI, paraphrase, sentiment/emoji/emotion, dialog domain/intent/slot, semantic parsing, NER, relation, syntax, coref, summarization, QA, relevance, qgen, ...

[Many langs]

[Previous work]

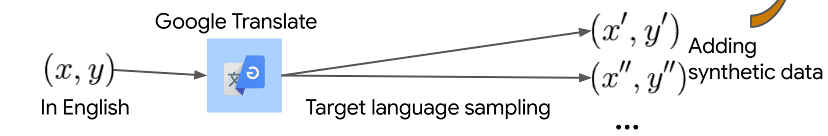
Fine-tuning a dense retriever to **find useful demonstrations** for ICL with LLMs



Contrastive learning (the encoder is **shared across tasks and languages**)

[Multi/cross-lingual data augmentation]

Translating some datasets **into many languages** before the utility estimation

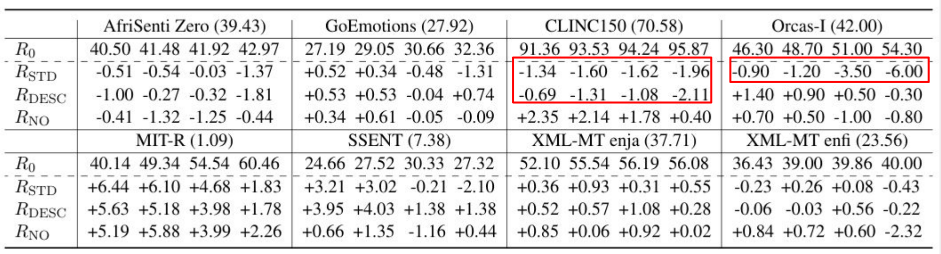


Research Question (1)

Q. Does the **demonstration text format** matter when tested on unseen tasks?

A. Yes, using (x, y) instead of (x) alone would hurt the generalization ability.

0: generic mT5 retriever **STD**: (instruct, x, y) **DESC**: (instruct, x, description(y)) **NO**: (instruct, x)



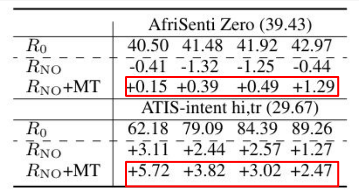
	AfriSenti Zero (39.43)				GoEmotions (27.92)				CLINC150 (70.58)				Orcas-I (42.00)			
R_0	40.50	41.48	41.92	42.97	27.19	29.05	30.66	32.36	91.36	93.53	94.24	95.87	46.30	48.70	51.00	54.30
R_{STD}	-0.51	-0.54	-0.03	-1.37	+0.52	+0.34	-0.48	-1.31	-1.34	-1.60	-1.62	-1.96	-0.90	-1.20	-3.50	-6.00
R_{DESC}	-1.00	-0.27	-0.32	-1.81	+0.53	+0.53	-0.04	+0.74	-0.69	-1.31	-1.08	-2.11	+1.40	+0.90	+0.50	-0.30
R_{NO}	-0.41	-1.32	-1.25	-0.44	+0.34	+0.61	-0.05	-0.09	+2.35	+2.14	+1.78	+0.40	+0.70	+0.50	-1.00	-0.80
	MIT-R (1.09)				SSENT (7.38)				XML-MT enja (37.71)				XML-MT enfi (23.56)			
R_0	40.14	49.34	54.54	60.46	24.66	27.52	30.33	27.32	52.10	55.54	56.19	56.08	36.43	39.00	39.86	40.00
R_{STD}	+6.44	+6.10	+4.68	+1.83	+3.21	+3.02	-0.21	-2.10	+0.36	+0.93	+0.31	+0.55	-0.23	+0.26	+0.08	-0.43
R_{DESC}	+5.63	+5.18	+3.98	+1.78	+3.95	+4.03	+1.38	+1.38	+0.52	+0.57	+1.08	+0.28	-0.06	-0.03	+0.56	-0.22
R_{NO}	+5.19	+5.88	+3.99	+2.26	+0.66	+1.35	-1.16	+0.44	+0.85	+0.06	+0.92	+0.02	+0.84	+0.72	+0.60	-2.32

k-shot ICL with k=1, 3, 5, 10 (**no data augmentation with MT**)

Research Question (2)

Q. Does MT help?

A. Yes.



	AfriSenti Zero (39.43)			
R_0	40.50	41.48	41.92	42.97
R_{NO}	-0.41	-1.32	-1.25	-0.44
R_{NO}+MT	+0.15	+0.39	+0.49	+1.29
	ATIS-intent hi,tr (29.67)			
R_0	62.18	79.09	84.39	89.26
R_{NO}	+3.11	+2.44	+2.57	+1.27
R_{NO}+MT	+5.72	+3.82	+3.02	+2.47

Other aspects are also investigated in our paper.

Discussions: challenges towards even better generalization

# Adaptability to arbitrary formats of tasks by real users

- We carefully selected and processed the datasets, but they are (too) clean.
- In real usecases of LLM APIs, the users will type their tasks in arbitrary (potentially more complicated) forms.

# Understanding of more nuanced task instructions (i.e., controllability)

- Can the retriever have capabilities of understanding the tasks in nuanced ways?
- For example, instead of just “machine translation,” we may have specific priorities like entity precision/recall, writing styles, etc.