Can You Answer This? Exploring Zero-Shot Question Answering Generalization Capabilities in Large Language Models

Saptarshi Sengupta *[1], Shreya Ghosh[1], Preslav Nakov[2], and Prasenjit Mitra[1]

[1]College of IST, The Pennsylvania State University, USA
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE

Abstract

The buzz around Transformer-based Language Models (TLM) is well founded owing to their impressive performance on a number of tasks. However, when applied to areas that require specialized knowledge (closed-domain), such as medical, finance, etc., their performance takes a drastic hit, sometimes more than their older recurrent/convolutional counterparts. In this paper, we explore the *zero-shot* capabilities of large Language Models (LMs) for extractive Question Answering (QA). In particular, we study the performance changes in the face of *domain drift*, i.e., when the target domain data is vastly different in terms of semantic and statistical properties from the source domain, and we attempt to explain the subsequent behavior. To this end, **we present two studies**. Our findings indicate certain flaws in the current generation of TLMs, which prevent them from performing well on closed-domain tasks.

1 Introduction

The optimism surrounding the achievements of Transformer-based Language Models (TLMs) such as BERT [3], RoBERTa [5], and XLNet [12] starts getting tested when we probe their generalization capabilities across domains. According to the established pre-training + fine-tuning framework, these models should be fine-tuned on every new dataset they are to be applied to in order to achieve state-of-the-art (SOTA) performance. However, an increasing number of studies [6, 8, 4] have started questioning the limits of this approach and exploring *no-fine-tune* or *zero-shot* and *few-shot* transfer learning settings.

*Corresponding Author: sks6765@psu.edu

Closed-domain datasets (CDD) pose challenges that TLMs are not equipped to tackle. First, the language used to describe phenomena in their space is quite dense i.e., filled with jargon that typically does not appear in general purpose corpora (an out-of-vocabulary issue). Second, the number of samples in these datasets is not nearly enough to provide competitive performance via fine-tuning, since in order to work well, TLMs need numerous samples to learn a decent enough approximation of the data. Third, the statistical demands of these datasets such as answer and context lengths are much more than what these TLMs are used to see.

The above properties, while useful for understanding the nuances of CDD, do not completely explain *why* there exist performance discrepancies in the case of domain shifts. For example, even if we use models specifically tailored for a given domain, say biomedical, we see that their performance is still quite poor. For example, SciBERT [1] when applied to COVID-QA [7] yields mediocre zero-shot performance of $\approx$ 0.45 F1 & 0.25 EM, and even when fine-tuned it does not improve significantly, 0.54 F1 & 0.29 EM [10].

Thus, here we investigate two hypotheses, based on a study of related datasets and literature, with teh aim to shed some light on the root cause for such performance discrepancies: (*i*) *generated answer length analysis* and (*ii*) *semantic similarity between the different senses of polysemous domain terms.* We have observed that CDD setups usually demand longer answer spans to be generated as opposed to simple factoid-based QA. This makes sense as we need detailed answers to questions such as *Why is remdesivir ineffective for COVID-19?* as opposed to simple factoid questions. As such, we wanted to check whether these models are capable of producing longer spans as predictions.

Word	Server			Java			Windows			following					
Model \Sense	host (s1)	waiter (s2)	(s1, s2)	software (s1)	coffee (s2)	(s1, s2)	software (s1)	framework (s2)	(s1, s2)	reference (s1)	pursue (s2)	(s1, s2)	Min	Max	Avg.
BERT	0.78	0.8	0.48	0.84	0.67	0.48	0.77	0.71	0.33	0.61	0.58	0.31	0.31	0.84	0.61
RoBERTa	0.94	0.95	0.89	0.94	0.91	0.84	0.93	0.94	0.79	0.91	0.92	0.84	0.79	0.95	0.90
SciBERT	0.79	0.82	0.71	0.79	0.65	0.59	0.71	0.85	0.62	0.67	0.69	0.55	0.55	0.85	0.70
SenseBERT	0.88	0.89	0.75	0.90	0.69	0.53	0.83	0.92	0.76	0.72	0.81	0.52	0.52	0.92	0.77

Table 1: Average semantic similarity between contextualized (s)enses of the domain terms in TechQA.

	SQuADv1			
Model	Gold	Predicted	EM	F1
BIDAF	18.73	**25.31**	65.73	75.98
QANet		23.74	26.3	36.81
BERT		18.18	80.95	88.25
RoBERTa		18.03	**82.73**	**90.04**

Table 2: Average number of characters in the answer spans for validation questions in SQuAD. **Zero-shot performance (EM & F1) is also shown**.

	TechQA			
Model	Gold	Predicted	EM	F1
BIDAF	156.79	**4302.93**	**32.23**	**39.45**
QANet		387.2	3.96	7.65
BERT		18.42	1.61	6.35
RoBERTa		26.89	1.94	4.68

Table 3: Average number of characters in the answer spans generated for questions in the validation set of TechQA.

The rationale for the second test is that, typically such datasets would showcase only one sense of a polysemous word, i.e., the one that is *usually* talked about in that space. Thus, we needed to see whether models can distinguish desired from auxiliary senses.

2 Answer Length Analysis

We computed the average number of characters present in the generated answer spans on the validation set of TechQA [2] (our cross-domain dataset, or CDD) and SQuAD [9] (our out-of-domain dataset, or ODD). We chose SQuAD as the ODD since there exist many fine-tuned models for it, which means we do not need to do retraining. We count characters instead of tokens as we want a scheme that is independent of a particular tokenization strategy (e.g., BPE, word-piece, etc.) We present our results in Tables 2 and 3.

We see that the CDD setup requires longer answer spans of text which the TLM are not able to generate. To see whether this is a problem with neural architectures in general, we examine scores from two other variants viz. recurrent –BIDAF [11]–, and convolutional –QANet [13]. While both are capable of generating incredibly longer spans of text, BIDAF performs multitudes better than either TLM in **this CDD** setup (remains to be seen for other CDD setups). Further analysis is needed to determine whether this behavior exists across other CDD setups and/or whether it higher scores are simply due to longer spans because the current F1 measure is based on token overlap.

3 Semantic Similarity Trials

We conjecture that polysemous words in CDD only manifest the *dominant* sense in their space. Given the number of samples, it would be difficult to expect a technical QA dataset to have (m)any instances of the *coffee* sense for *Java* (although this may not always be the case as in *cold* for temperature & condition for biomedical datasets). To test this hypothesis, we create a dataset of polysemous domain terms, appearing in the vocabulary of the TLM and the corpus (TechQA), and associated contexts. As expected, the corpus shows only a single sense of a word and as such, we scraped an online resource, `vocabulary.com`, for contexts for the other sense. In total, we had ten contexts per sense of a given word. We compute average cosine similarity b/w contextualized embeddings of the target word from same and different sense groups. The logic here is that intra/same-group similarity should be high while inter/different-group should be low. According to Table 1, BERT *is able* to respect this logic whereas the other methods, including SenseBERT (surprisingly) cannot.

4 Conclusion

Acknowledging these results are preliminary, we feel that scores from the presented trials do provide some insight into the poor performance of TLMs in a CDD setup. However, we need to run these experiments and more on other datasets to make our claims ironclad.

References

[1] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

[2] V. Castelli, R. Chakravarti, S. Dana, A. Ferritto, R. Florian, M. Franz, D. Garg, D. Khandelwal, J. S. McCarley, M. McCawley, et al. The techqa dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1269–1278, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[4] A. Kumar and V. H. C. Albuquerque. Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20 (5):1–13, 2021.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Q. Lyu, H. Zhang, E. Sulem, and D. Roth. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.42. URL https://aclanthology.org/2021.acl-short.42.

[7] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.18.

[8] M. Moradi, K. Blagec, F. Haberl, and M. Samwald. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*, 2021.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[10] S. Sengupta, C. Heaton, S. Sarkar, and P. Mitra. Leveraging external knowledge resources to enable domain-specific comprehension. Accepted to CoLLAs Workshop, 2022.

[11] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[13] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.