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

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

[1] College of IST, The Pennsylvania State University, USA
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{sks6765, shreya, pmitra}@psu.edu, preslav.nakov@gmail.com

Abstract

The buzz around Transformer-based language models (TLM) such as BERT, RoBERTa, etc. is well founded owing to their impressive results on an array of tasks. However, when applied to areas needing specialized knowledge (closed-domain), such as medical, finance, etc. their performance takes drastic hits, sometimes more than their older recurrent/convolutional counterparts. In this paper we explore *zero-shot* capabilities of large LMs for extractive QA. Our objective is to examine performance change in the face of *domain drift* i.e. when the target domain data is vastly different in semantic and statistical properties from the source domain and attempt to explain the subsequent behavior. To this end, **we present two studies** in this paper while planning further experiments later down the road. Our findings indicate flaws in the current generation of TLM limiting their performance on closed-domain tasks.

Introduction

The optimism surrounding the BERT (Rogers, Kovaleva, and Rumshisky 2020) & related family of TLM gets tested when probing their generalization capabilities i.e. capacity to perform well across domains. According to the established pre-training + fine-tuning framework, these models should be fine-tuned on every new dataset it is to be applied on to achieve SOTA (or high) performance. However, an increasing number of studies (Lyu et al. 2021; Moradi et al. 2021) are questioning the limits of this approach and exploring *zero/few-shot* transfer settings.

Closed-domain datasets (CDD) present challenges which TLM (current generation) are not equipped for tackling. First, the language used to describe phenomena in their space is quite dense i.e. filled with jargon rarely appearing in general purpose corpora (out-of-vocabulary issue). Second, the number of samples in these datasets is insufficient to command competitive performance via fine-tuning since to work well, TLM need numerous samples to learn a decent approximation of the data. Third, the statistical demands of these datasets such as answer and context lengths are much more than what these TLM are used to seeing.

While the above properties are useful for understanding the nuances of CDD, they are inadequate in explaining the performance gap due to domain shift. For example, even if we use models specifically tailored for a domain, say biomedical, we observe their performance is still poor; ex. SciBERT (Beltagy, Lo, and Cohan 2019) when applied to COVID-QA (Möller et al. 2020) gives mediocre performance (Sengupta et al. 2022), $\approx$ 0.45 F1 & 0.25 EM, and even when fine-tuned does not improve significantly, 0.54 F1 & 0.29 EM.

Motivated by the challenges, we seek to investigate two hypotheses based on a study of related datasets and literature to understand the root cause for the same. We have observed that CDD usually demand longer answer spans to be generated. We need detailed answers to questions such as *Why is remdesivir ineffective at treating COVID-19?* as opposed to short answers provided in response to simple factoid-based QA. Therefore, we ask whether these models are capable of identifying and producing longer spans of context (‡1)? Next, **we conjecture** that polysemous words in CDD only manifest the *dominant* sense in their space. Ex. 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). Thus, we wanted to see if TLM are capable of distinguishing the desired sense from the auxiliary ones (‡2). An inability to do so would indicate a deep semantic drawback.

Answer Length Analysis (‡1)

We recorded the average number of characters in the generated answer spans on the validation set of TechQA (Castelli et al. 2019) (our CDD) and SQuAD (Rajpurkar et al. 2016) (our ODD). SQuAD was chosen as the ODD since there exist many pre-fine-tuned models, eliminating the need for retraining. We count characters instead of tokens since we wanted to use a scheme consistent across all models irrespective of their tokenization strategy (ex. BPE, wordpiece, etc.)

Table 2 indicates that CDD require longer answer spans of text which the TLM are unable to generate. To see whether this is a problem with neural architectures in general, we examine scores from two other variants, recurrent [BIDAF (Seo et al. 2016)] and convolution [QANet (Yu et al. 2018)]. While both are capable of generating incredibly longer spans of text, irrespective of the domain, interestingly, BIDAF performs significantly better than either TLM on **this CDD**. Further analysis is needed to determine whether this behavior exists across other CDD and/or whether it achieves high scores sim-

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: Avg. Semantic Similarity between contextualized (s)enses of the domain terms as found in TechQA.

ply due to its longer spans because the current F1 metric is based on token overlap. We would be circumspect to linearly correlate performance with predicted sequence length.

	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 no. of characters in the answer spans generated for questions in the validation set of SQuAD. **Note, zero-shot performance (EM, F1) is also provided alongside.**

	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 no. of characters in the answer spans generated for questions in the validation set of TechQA.

Semantic Similarity Trials (‡2)

We created a dataset of polysemous domain terms, appearing in the vocabulary of the TLM and TechQA, and associated contexts. As expected, the corpus shows only a single sense of a word. We scraped vocabulary.com for contexts for the other sense of the words. 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 & different sense groups. We expect that intra/same-group similarity should be high while inter/different-group similarity should be low or at least the margin should be large enough to indicate the models' ability to segregate senses.

The following can be said about each model according to Table 3. BERT shows the most range of values across the board, rarely breaking the 0.8 mark, in line with Ethayarajh (2019), and in turn achieves the lowest average similarity (due to extremes). This could indicate a striking ability of BERT to distinguish senses. Despite RoBERTa's higher performance (Table 1) it consistently gives higher cosine similarity, which is unexpected & could indicate that its representations are densely packed in its embeddings space. If this is indeed true, we need further insight into how it distributes its data points which in turn could lead to an understanding of the geometry of BPE v/s word-piece embeddings. SciBERT oddly seems to favor auxiliary senses over relevant ones (only *Java's* (*Software*) sense is higher than (*coffee*)). Finally, SenseBERT (Levine et al. 2019), as expected, supports our hypothesis to an extent. This could indicate a necessity for *fusing external information* during pre-training.

Conclusion

The presented trials demonstrate preliminary but interesting insights into poor performances of TLM on CDD. We plan to extend these experiments & test on other datasets to strengthen the fundamentals of our claims.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Castelli, V.; Chakravarti, R.; Dana, S.; Ferritto, A.; Florian, R.; Franz, M.; Garg, D.; Khandelwal, D.; McCarley, S.; McCawley, M.; et al. 2019. The techqa dataset. *arXiv preprint arXiv:1911.02984*.

Ethayarajh, K. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Levine, Y.; Lenz, B.; Dagan, O.; Ram, O.; Padnos, D.; Sharir, O.; Shalev-Shwartz, S.; Shashua, A.; and Shoham, Y. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Lyu, Q.; Zhang, H.; Sulem, E.; and Roth, D. 2021. Zero-shot Event Extraction via Transfer Learning: Challenges and Insights. In *ACL/IJCNLP (2)*.

Möller, T.; Reina, A.; Jayakumar, R.; and Pietsch, M. 2020. COVID-QA: a question answering dataset for COVID-19.

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

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8: 842–866.

Sengupta, S.; Heaton, C.; Sarkar, S.; and Mitra, P. 2022. Leveraging External Knowledge Resources to Enable Domain-Specific Comprehension. Accepted to CoLLAs Workshop.

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

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