Typologically Diverse QA - Zero-Shot and Few Shot Language Jackknifing

Kumar Tanmay
Department of Electrical Engineering
IIT Kharagpur
kr.tanmay147@iitkgp.ac.in

Abstract

Natural Language Understanding has seen tremendous progress across various tasks such as machine translation, QA, and natural language inference. However, most of the progress is limited to English and other high-resource languages where parallel data is easily available. Hence, recently there has been a shift of focus to cross-lingual generalizing models. Particularly, multilingual BERT has been shown to perform cross-lingual generalization quite well [1]. But it has been observed that this cross-lingual generalization depends quite on the typological similarities between the train and test languages and performance degrades when evaluation is performed on languages with a typology different than the languages on which mBERT is trained [1]. In this project, we utilize the TYDI QA dataset [2] to explore the cross-lingual generalization ability of mBERT for the QA task and study the effect of different linguistic features of the languages. Concretely, we study the performance of mBERT over QA in zero-shot settings, where we train on a subset of languages and test on the rest and analyze the mistakes committed. We observe that interpretation of questions in most cases in incorrect. To understand this, we also fine-tune mBERT as a language model over questions from the TYDI QA dataset and repeat the zero-shot experiments. We observe that zero-shot transfer improves because of language modeling fine-tuning task.

1 Approach

Methodology: In this work, we utilize the TYDI QA dataset [2] to explore the cross-lingual generalization ability of mBERT and study the effect of different linguistic features of the languages. We evaluate the performance of mBERT over the task of question-answering in the regime of zero data across different languages. We study zero-shot QA by explicitly not training on a subset of the provided languages. Particularly, we fine-tune mBERT on all languages of TYDI QA dataset except one, and evaluate its performance on the devset of the left-out (target) language. We also fine-tune mBERT as a language model over questions from the TYDI QA dataset and repeat the zero-shot experiments.

Baselines: We use the following baselines: (a) Using a pre-trained mBERT and measuring its performance on the full TYDI QA gold passage dev set, (b) fine-tuning mBERT on the 100k English-only SQuAD 1.1 trainset and evaluating on the dev set, (c) fine-tuning mBERT on the SQuAD 1.1 trainset followed by fine-tuning over the trainset of all the languages in the TYDI QA gold passage dataset and evaluating on the dev set, (d) fine-tuning mBERT over the train set of all the languages in the TYDI QA gold passage dataset and measuring its performance over the dev set, and (e) fine-tuning mBERT separately for each of the different languages, and then evaluating on the corresponding language's dev set.

Stanford CS224N Natural Language Processing with Deep Learning

We use an existing implementation of mBERT in PyTorch (from HuggingFace) for all our experiments.

2 Experiments

Data: The TYDI QA [2] dataset consists of a total of 204K question-answer pairs across 11 typologically different languages. However, we use the secondary Gold Passage task dataset (TYDI QA-GoldP) for our study as it is more similar to existing simple reading comprehension datasets, where only the gold answer passage is provided and unanswerable questions have been discarded. This setting allows us to study the how the typological features of different languages and mBERT interplay without other confounding variables. Table 1 shows the data statistics of the TYDI QA-GoldP dataset (consisting of 9 languages).

Evaluation: We use two different metrics to quantitavely evaluate model performance: (a) Exact Match (EM): measures the percentage of predictions that match any one of the ground truth answers exactly, and (b) F1 score: measures the average overlap between the prediction and ground truth answer (harmonic mean of precision and recall). Both are commonly used metrics for QA in literature [3, 2].

Results: Table 3 shows the baseline results of fine-tuning mBERT seperately for each language from TYDI QA GoldP dataset. Table 4 shows the results in zero-shot setting for each target language. The last row in Table 3 corresponds to fine-tuning mBERT on all the languages. We observe an expected trend in the results. For each language, the model performance improves from zero-shot setting to fine-tuning on only that language to fine-tuning on all the languages. Table 2 contains the other baseline results.

Analyzing Failure cases of Zero-shot: We analyze the questions that are correctly answered when we fine-tune mBERT on all languages vs when the language under consideration is left out during training (zero-shot). We find that in most cases the interpretation of the questions across languages is a little off, for example, when we leave out Swahili during training and evaluate on its devset, the model predicted *place of birth* when the question was based on *date of birth* (refer Table 6 for more such examples).

Discussion: Based on the observations, we hypothesize that the weak zero-shot transfer performance can be caused due to the following reasons: (a) mBERT might not have seen a lot of question words in different languages as it is trained on Wikipedia data, (b) The model does not have a mapping of question words to the type of answer to predict. To test the first hypothesis we carry out a simple experiment where we finetune mBERT as a Language Model with just the questions from the TYDI QA GoldP dataset and then repeat the zero-shot experiments, the results of which are given in Table 5. We find that for all the languages, except Swahili and Telugu, the zero-shot transfer improves because of the language modeling task. This improvement is particularly significant ($\sim$ 7 points) for few languages like Finnish and Indonesian. However, we find that there is a decrease in performance when the model is trained on all languages (last row in Table 5).

Experimental Details: For all the experiments where we finetune mBERT for the QA task, we use the same hyperparameter configurations: Learning Rate = 3e-5, Epochs = 2, Per-gpu batch-size = 5. To finetune for the language modelling task, we follow the standard protocol where a token is selected for masking with a probability of 0.15 and the token is replaced by (a) a [MASK] token 80% of the time, (b) a random word 10% of the time and, (c) the same word 10% of the time. The model is finetuned for 1 epoch with a learning rate of 5e-5 and per-gpu batch size of 4.

3 Future work

The rest of the project entails experiments in the direction of few-shot learning and doing a deeper analysis of the incorrect predictions. We intend to further analyze the patterns of zero-shot errors and find if other such patterns exist in different languages. In addition to

the LM experiment, we plan to conduct further experiments to examine our hypothesis for the reason behind the errors. One such experiment is to replace the question word of the trainset languages with that of the left-out language. Another experiment is to construct a question-type classification task and see if the zero-shot performance on this task correlates with that of zero-shot QA and if pretraining mBERT with question-type classification task helps in zero-shot performance on QA.

We also intend to explore the extent of cross-lingual generalization by training on a relatively small amount of new data of a target language and monitoring changes in the task performance as we increase training data for that language. Hopefully this would provide some insights into the effect of amount of data required to achieve better transferability. We will also explore fine-tuning a smaller model, DistillmBERT [4], to see the effect of model size.

Language	**Finnish**	**Telugu**	**Bengali**	**Arabic**	**Indonesian**	**Korean**	**Russian**	**Swahili**	**English**
Train	7249	8171	3503	17849	5953	2041	7706	3157	3862
Dev	844	669	113	951	598	308	830	518	495

Table 1: Data Statistics of TYDI QA GoldP task.

Language	**No training**	**Trained on SQuAD**	**Trained on SQuAD + TyDiQA**	**Trained on TyDiQA**
Arabic	0.11 (9.33)	50.89 (66.57)	72.98 (84.55)	74.45 (85.20)
Bengali	0.00 (3.41)	49.56 (62.35)	66.37 (77.34)	69.03 (80.31)
English	0.20 (7.60)	69.70 (78.78)	73.54 (83.45)	71.31 (81.74)
Finnish	0.24 (7.57)	45.50 (57.35)	71.21 (80.97)	70.26 (81.17)
Indonesian	0.17 (8.29)	49.83 (63.09)	75.25 (85.03)	75.92 (85.30)
Korean	0.00 (3.90)	46.75 (53.88)	63.31 (72.54)	62.66 (71.06)
Russian	0.24 (5.93)	46.63 (65.13)	69.16 (80.93)	69.40 (80.35)
Swahili	0.00 (6.70)	47.30 (60.20)	80.69 (87.26)	79.73 (85.93)
Telugu	0.00 (2.29)	44.84 (52.68)	71.00 (84.46)	71.15 (84.26)

Table 2: Experimental results for different scenarios: (a) **No training**: Use pre-trained mBERT and directly evaluate on the devset, (b) **Trained on SQuAD**: Fine-tune mBERT using SQuAD, (c) **Trained on SQuAD+TYDI QA**: Fine-tune mBERT on both SQuAD and TYDI QA dataset, (d) **Trained on TYDI QA**: Fine-tune mBERT on whole TYDIQA dataset.

Trained Language	**Arabic**	**Bengali**	**English**	**Finnish**	**Indonesian**	**Korean**	**Russian**	**Swahili**	**Telugu**
Arabic	**73.08(84.83)**	44.25(56.52)	54.55(66.63)	47.75(63.41)	53.51(68.94)	50.97(58.81)	43.25(65.83)	48.26(63.49)	45.29(52.45)
Bengali	24.08(44.99)	**61.06(74.87)**	37.37(54.65)	32.70(51.56)	35.12(51.66)	47.40(55.72)	31.08(50.51)	44.59(64.89)	41.70(52.67)
English	49.11(66.00)	38.05(54.70)	**64.44(75.29)**	47.63(60.92)	54.52(68.61)	47.08(55.82)	42.17(59.82)	38.03(59.92)	39.61(49.55)
Finnish	52.47(67.65)	45.13(61.64)	51.31(63.20)	**68.48(79.79)**	50.00(66.13)	46.10(56.43)	40.12(62.74)	47.30(63.48)	43.35(53.76)
Indonesian	55.73(70.14)	46.02(59.70)	61.62(72.41)	49.05(64.20)	**73.91(84.04)**	49.35(58.49)	44.70(62.86)	45.95(62.79)	41.55(50.85)
Korean	32.70(49.65)	46.90(58.55)	46.06(58.60)	36.97(50.80)	38.63(52.66)	**57.79(66.70)**	32.65(50.22)	45.17(60.41)	40.06(51.74)
Russian	53.63(70.67)	40.71(55.77)	60.40(73.16)	47.87(62.30)	52.68(68.71)	45.78(56.54)	**66.51(77.65)**	42.86(61.23)	45.89(54.11)
Swahili	36.38(53.50)	40.71(52.48)	38.99(51.38)	38.15(52.23)	42.81(56.12)	43.83(54.14)	26.75(44.45)	**74.52(81.69)**	37.97(46.40)
Telugu	32.28(47.73)	51.33(63.96)	47.88(60.66)	41.71(53.31)	40.30(51.26)	47.40(57.40)	43.73(56.98)	47.10(64.72)	**68.31(82.03)**
All	**74.45(85.20)**	**69.03(80.31)**	**71.31(81.74)**	**70.26(81.17)**	**75.92(85.30)**	**62.66(71.06)**	**69.40(80.35)**	**79.73(85.93)**	**71.15(84.26)**

Table 3: Results of fine-tuning mBERT only on one language (shown in rows) from TYDI QA GoldP dataset and evaluating its performance on all languages (shown in columns). For example, we fine-tune mBERT only on Arabic and evaluate its performance on the dev set of all the languages. The last row corresponds to fine-tuning mBERT on all the languages. Here, we report the Exact Match score and F1 score (inside brackets).

References

[1] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019. 1

Left out language	Arabic	Bengali	English	Finnish	Indonesian	Korean	Russian	Swahili	Telugu
Arabic	**58.15(72.30)**	64.60(76.76)	66.67(78.17)	69.55(80.54)	75.42(85.25)	58.44(67.84)	66.02(78.83)	79.15(85.32)	69.81(83.46)
Bengali	71.71(83.78)	**49.56(63.14)**	70.30(80.38)	69.55(80.66)	73.08(84.36)	62.34(70.58)	66.87(77.83)	80.12(85.93)	70.55(84.36)
English	72.77(83.91)	66.37(76.03)	**67.47(77.62)**	69.91(79.92)	76.09(85.60)	61.36(69.64)	66.87(77.94)	78.76(85.90)	70.25(83.70)
Finnish	72.87(84.20)	58.41(72.84)	69.29(79.92)	**53.79(66.35)**	76.09(85.08)	60.06(68.37)	67.83(78.60)	76.45(83.96)	70.10(83.40)
Indonesian	72.56(83.79)	60.18(73.15)	68.28(79.71)	69.67(80.50)	**57.02(71.02)**	60.39(68.52)	68.67(79.44)	77.22(84.82)	71.15(83.89)
Korean	73.19(84.51)	61.06(75.09)	69.49(78.91)	69.79(80.13)	74.41(85.13)	**54.55(63.33)**	67.71(79.34)	80.31(86.06)	69.66(83.56)
Russian	73.61(84.78)	65.49(77.63)	69.09(78.67)	70.02(80.23)	75.25(84.88)	63.64(71.76)	**47.59(71.11)**	79.34(85.21)	71.45(84.41)
Swahili	72.45(83.92)	65.49(77.73)	68.28(79.34)	68.60(79.01)	73.91(84.46)	61.36(70.19)	67.95(79.27)	**53.47(69.25)**	70.40(83.99)
Telugu	72.87(84.11)	63.72(74.95)	68.08(78.11)	70.50(80.75)	73.91(84.42)	61.04(70.40)	68.19(79.42)	77.41(85.67)	**47.09(56.39)**

Table 4: **Zero Shot Setting**. Results of fine-tuning mBERT on all languages except one (shown in rows) from TYDI QA GoldP dataset and evaluating on all languges (shown in columns). For example, first row corresponds to fine-tuning mBERT on all but Arabic and its performance on the dev set of all the languages. Here, we report the Exact Match score and F1 score (inside brackets).

Left out language	Arabic	Bengali	English	Finnish	Indonesian	Korean	Russian	Swahili	Telugu
Arabic	**61.69 (75.62)**	61.54 (74.25)	68.14 (77.75)	65.82 (76.90)	73.56 (83.06)	57.75 (67.69)	65.55 (76.61)	77.72 (84.19)	70.00 (83.16)
Bengali	71.25 (82.34)	**50.77 (61.21)**	67.67 (78.34)	66.84 (77.49)	73.14 (83.29)	59.63 (68.26)	65.34 (76.67)	76.19 (82.94)	71.57 (84.02)
English	70.52 (82.81)	60.00 (71.33)	**66.72 (77.42)**	67.96 (77.85)	71.73 (83.10)	61.76 (70.43)	65.34 (77.13)	75.68 (82.56)	71.14 (83.81)
Finnish	70.25 (82.51)	66.92 (79.75)	67.82 (78.23)	**60.83 (72.54)**	72.57 (82.72)	59.63 (68.26)	67.23 (78.45)	76.36 (83.01)	71.00 (83.73)
Indonesian	70.79 (82.78)	60.77 (73.21)	67.67 (78.55)	66.63 (77.07)	**64.28 (78.34)**	60.70 (69.44)	65.76 (77.17)	77.55 (84.30)	71.00 (83.65)
Korean	69.15 (81.40)	65.38 (75.51)	66.88 (77.26)	67.85 (78.66)	73.00 (82.78)	**58.56 (65.81)**	65.65 (76.68)	77.89 (83.71)	71.14 (83.88)
Russian	71.25 (82.96)	62.31 (74.70)	66.88 (78.36)	65.51 (76.68)	72.86 (83.03)	61.76 (69.47)	**50.05 (70.47)**	75.85 (82.70)	71.00 (83.71)
Swahili	68.97 (81.57)	58.46 (71.60)	65.62 (77.34)	65.51 (76.93)	72.01 (82.89)	59.36 (68.48)	65.65 (76.90)	**49.49 (63.79)**	70.57 (83.15)
Telugu	70.25 (82.48)	64.62 (75.55)	66.40 (77.41)	67.04 (77.81)	73.28 (83.36)	61.23 (68.87)	65.34 (77.07)	76.02 (81.52)	**45.57 (52.79)**
All	**70.06 (82.74)**	**67.69 (79.22)**	**67.35 (78.00)**	**66.33 (76.84)**	**72.57 (82.23)**	**59.36 (68.50)**	**65.03 (76.72)**	**77.55 (83.23)**	**70.43 (83.72)**

Table 5: Here, we fine-tune mBERT as a Language Model with just the questions from the TYDI QA GoldP dataset and repeat the zero-shot experiments. We hypothesize that this will allow flexibility to the model in interpreting questions of different languages.

[2] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 2020. 1, 2

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

Question	GT Answer	ZeroShot Answer
	Bengali	
দ্বিতীয় এশিয়া কাপ ক্রিকেট টুর্নামেন্ট কোথায় অনুষ্ঠিত হয়া ?	শ্রীলঙ্কা	১৯৮৬
Dbitīya ēśiẏā kāpa krikēṭa ṭurnāmēnṭa kōthāẏa anuṣṭhita haẏa?	Śrīlaṅkā	
Where was the Second Asia Cup Cricket Tournament held?	Sri Lanka	1986
বাংলাদেশের উচ্চ আদালত মোট কতগুলি ভাগে বিভক্ত ?	দুটি	আপীল বিভাগ এবং হাইকোর্ট বিভাগ
Bānlādēśēra ucca ādālata mōṭa kataguli bhāgē bibhakta?	Duṭi	Āpīla bibhāga ēbaṁ hā'ikōrṭa bibhāga
How many section does the Supreme Court of Bangladesh have?	Two	Appellate Division and High Court Division
ভ্লাদিমির ইলিচ উলিয়ানভ ওরফে লেনিনের জন্ম কবে হয় ?	১৮৭০ সালে ২২শে এপ্রিল	সিমবির্স্ক
Bhlādimira ilica uliẏānabha ōraphē lēninēra janma kabē haẏa?	1870 Sālē 22śē ēprila	Simabirska
When was Vladimir Ilych Ulyanov alias Lenin born?	April 22, 1870	Simbirsk
	Swahili	
Je,Howard Winchel Koch alianza utayarishaji filamu lini?	1947	Universal Studios
When did Howard Winchel Koch start filming?		
Ayatollah Ruhollah Musawi Khomeini alizaliwa wapi?	Khomein	24 Septemba 1902
Where was Ayatollah Ruhollah Musawi Khomeini born?		September 24, 1902
Mtume Mohammed alikuwa na wake wangapi?	Three	Aisha bint Abu Bakr
How many wives did the Prophet Mohammed have?		Aisha, daughter of Abu Bakr
	Finnish	
Milloin Richard Rorty syntyi?	4. lokakuuta 1931	New York, New York, Yhdysvallat
When was Richard Rorty born?	October 4, 1931	New York, New York, United States
Missä Mika Waltar kuoli?	Helsinki, Suomi	26. elokuuta 1979
Where did Mika Waltar die?	Helsinki, Finland	August 26, 1979
Kuinka korkea on Kalliovuorten korkein huippu?	4399 metriä	Mount Elbert Coloradossa
How high is the highest peak in the Rockies?	4399 meters	Mount Elbert in Colorado

Table 6: Example Predictions in zero-shot setting where the model predicts totally incorrect answer compared to when trained on all the languages. The transliteration (where the script is not Latin) and translation is provided for each example below it. We can see that there exists a pattern of mistakes across languages in the zero-shot scenario, for e.g., the model predicts *location* ("where") for *time* based ("when") questions (e.g. "Universal Studios" instead of "1947" in Swahili, "Simbirsk" instead of "April 22, 1870" in Bengali). Similarly, the model predicts "which" for *attribute* based questions ("how many", "how high", etc) (e.g. "Mount Elbert Coloradossa" ("highest mountain") instead of "4399 metriä" ("height of the hightest mountain") in Finnish, "Appellate Division and High Court Division" ("which division") instead of "Two" ("how many") in Bengali).