You Need Only Uncertain Answers: Data Efficient Multilingual Question Answering

Zhihao Lyu [1] **Danier Duolikun** [1] **Bowei Dai** [1] **Yuan Yao** [1] **Pasquale Minervini** [1] **Tim Z. Xiao** [1] **Yarin Gal** [2]

Abstract

Data scarcity is a major barrier for multilingual question answering: current systems work well with languages such as English where data is affluent, but face challenges with small corpora. As data labelling is expensive, previous works have resorted to pre-tuning systems on larger multilingual corpora followed by fine-tuning on the smaller ones. Instead of curating and labelling large corpora, we demonstrate a data efficient multi-lingual question answering system which only selects uncertain questions for labelling, reducing labelling efforts and costs. To realise this *Bayesian active learning* framework, we develop methodology to quantify uncertainty in several state-of-art attention-based Transfer question answering models. We then propose an uncertainty measure based on the variance of BLEU scores, and computed via Monte Carlo Dropout, to detect out-of-distribution questions. We finish by showing the effectiveness of our uncertainty measures in various out-of-distribution question answering settings.

1. Introduction

Question Answering (QA) is a central task in Natural Language Processing (NLP), and has attracted tremendous attention from researchers in different fields such as statistics (Berger et al., 2000) and physics (Abdi et al., 2018). While there has been significant improvement of QA systems in rich-resource languages such as English (Zhang et al., 2020; Lan et al., 2019), the development of QA in low-resource languages remain scarce, mainly due to data scarcity and the highly effort-demanding process of labelling QA data in such languages. Due to such barriers, *multilingual transfer learning* appears as a feasible approach to reducing the amount of data needed in the low-resource target language. In this approach, one fine-tunes a QA model that is pre-trained in a rich-resource language using a smaller dataset in target languages. However, it is still desirable that the amount of data needed in the target language is minimised.

Active learning methods are designed to reduce the amount of labelled data needed by only labelling the data that is considered more informative by the model. Therefore, The key to the performance of an active learning algorithm is the mechanism of measuring how informative a data instance is. Most of the existing works resort to seeing the data considered more *uncertain* by the model as being more informative. While various data uncertainty measurements have been developed in lots of areas (Liu et al., 2020; Yan et al., 2016), to the best of our knowledge the corresponding uncertainty meausure in QA remains scarce. In this work, we propose a modified Monte Carlo Dropout BLEU (a machine translation evaluation metric) variance uncertainty metric and demonstrate a data-efficient active learning strategy based on our proposed uncertainty metric. Such active learning approach is deployed in the multilingual transfer learning process to reduce the amount of QA data needed in the target language.

To justify our proposed metric, we use it to detect out-of-distribution samples, assuming the same language samples from the same distribution. Concretely, let input context-question pair $x \in \mathcal{X}$ and an answer $y \in \mathcal{Y}$ be random variables that follow a joint data distribution $P_{\text{in}}(x, y)$ (e.g. English data distribution). In addition, let $P_{\text{out}}(x, y)$ denote a data distribution that is different from $P_{\text{in}}(x, y)$ (e.g. a German data distribution). Assuming that we have trained our model on a dataset drawn from $P_{\text{in}}(x, y)$, given a new sample x', we want to determine whether x' is from $P_{\text{in}}(x)$ or $P_{\text{out}}(x)$. This can be done by evaluating the uncertainty of the trained model on x', and we find that our proposed uncertainty measure can clearly detect out-of-distribution data by assigning high uncertainties to samples from languages that are different from the training language.

The existing body of research on QA suggests that the Transformer (Vaswani et al., 2017) based pre-training methods

[1]Department of Computer science, University College London, London, UK [2]Department of Computer science, University of Oxford, Oxford, UK. Correspondence to: Zhihao Lyu <ucabzly@ucl.ac.uk>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2019) are effective for improving model accuracy and robustness. Therefore, we study the properties of our proposed uncertainty measure on state-of-the-art Transformer-based transfer learning QA models.

To the best of our knowledge, this is the first attempt to study Dropout-based uncertainty measure and the corresponding active learning application for QA. Precisely, we make the following contributions:

1. We propose several uncertainty metrics and use them to detect the out-of-training-distribution data.
2. We investigate the latent factors that have effects on uncertainty.
3. We integrate the uncertainty metrics we proposed as a sample selection strategy with active learning, and find our metric brings about 10% data efficiency improvement.

2. Background and Related work

Our work builds on top of (Gal & Ghahramani, 2016; Xiao et al., 2020; Liu et al., 2020). Although we mentioned in the preceding argument that there are many barriers that prevent us from obtaining uncertainty, instead of using the traditional Bayesian approaches to calculate model uncertainty which comes with a prohibitive computational cost, Gal & Ghahramani (2016) developed a new theoretical framework that uses Dropout (Srivastava et al., 2014) training in deep neural networks as an approximation of Bayesian inference in deep Gaussian processes, which is easily deployed and leads to practical uncertainty estimation.

Xiao et al. (2020) developed a measure of uncertainty designed specifically for long sequences of discrete random variables to detect out-of-training-distribution sentences in Neural Machine Translation, solving a major intractability in the naive application of existing approaches on long sentences. They analysed the BLEU score (Papineni et al., 2002), a method for the automatic evaluation of machine translation models, and concluded that the BLEU score variance:

$$\mathbb{E}_{\theta}\mathbb{E}_{y,y'\sim p_{\theta}(y|x)}\left[1-\text{BLEU}(y,y')\right]^2 \tag{1}$$

produces a substantial improvement in separating in-distribution and out-of-distribution sentences. In Eq. (1), p denotes the translation model, θ denotes the model parameters, x denotes the input and y, y' denote different output samples respectively. In this work, we generalise the approach in Xiao et al. (2020) to QA, and propose some modifications to make the uncertainty metric work well on the short sentence.

There is a large body of research on active learning, and Liu et al. (2020) proposes an uncertainty-based active learning strategy called Lowest Token Probability (LTP) on BERT-CRF model and shows that the this method performs better than other strategies on both token-level F_1 and sentence-level accuracy. Our work is similar to this but instead of obtaining uncertainty by conditional random field, we directly use the sample uncertainty generated by MC Dropout.

3. Methods

In this section we first introduce the uncertainty metrics we intend to investigate, providing theoretical background and implementation details, then our investigation approach is described at high-level.

3.1. Uncertainty Metrics

3.1.1. Deterministic Sequence Probability

For each context-question pair x we assign uncertainty by the probability of chosen y (answer), the answer generated by a BERT-based model given the input x as follows.

$$\text{DSP} = \log\left[p(y|x)\right] \tag{2}$$

This metric is deterministic because once we have trained a model, the corresponding probability of chosen y with given x is determined (in the evaluation stage). Unlike in machine translation case as mentioned in Xiao et al. (2020), we do not penalise the sequence length because here we find the correct answer span that only need to calculate the probability of start and end position.

3.1.2. MC-Dropout Sequence Probability

We assign uncertainty to input x by:

$$\text{MCSP} = \log \mathbb{E}_{\theta\sim q(\theta)}\left[p_{\theta}(y|x)\right] \tag{3}$$

where q denotes the distribution of parameters. But in practice, we use the Monte Carlo Dropout and take N iterations to estimate the corresponding expectation.

$$\text{MCSP} = \log\left(\frac{\sum_{i=1}^{N} p_{\theta_i}(y|x)}{N}\right) \tag{4}$$

3.1.3. MC-Dropout Sequence Probability Variance

For this metric, instead of calculating the expectation value of MC sequence probability, we calculate their variance.

$$\text{MC-VAR} = \log\left[\mathbb{V}_{\theta\sim q(\theta)} p_{\theta}(y|x)\right] \tag{5}$$

Table 1. F1 of our trained baseline models/models trained in the MLQA paper on MLQA test set. For BERT, we used BERT-base while the authors (Lewis et al., 2019) trained BERT-large.

	en	de	zh
m-BERT	77.4/77.7	58.1/57.9	57.7/57.5
BERT	75.3/80.0	-/-	-/-
ALBERT	69.4/-	-/-	-/-

3.1.4. MC-DROPOUT BLEU SCORE VARIANCE

BLEU is the geometric mean of n-gram precision that is scaled by a brevity penalty to prevent very short sentences with little matching words from being given inappropriately high scores, and the standard BLEU score used for machine translation evaluation (BLEU: 4) is only really meaningful at the corpus level, since any sentence that does not have at least one 4-gram match will be given a score of 0. Since in QA there are many short answers, we propose a modified metric of MC-BLEU-VAR:

$$\mathbb{E}_\theta \mathbb{E}_{y,y' \sim p_\theta(y|x)} \left[1 - \text{BLEU}(y, y')\right]^2 \tag{6}$$

as

$$\text{MC-BLEU-VAR} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - \text{MBLEU}_{ij})^2 \tag{7}$$

Where MBLEU_{ij} denotes modified-BLEU and is defined as follows.

$$\begin{cases} \text{B}(\text{d}_{\theta_i}(x), \text{d}_{\theta_j}(x)), & l(\text{d}_{\theta_i}(x)) \geq 4 \text{ or } l(\text{d}_{\theta_j}(x)) \geq 4 \\ \text{F}_1(\text{d}_{\theta_i}(x),\ \text{d}_{\theta_j}(x)), & \text{otherwise} \end{cases} \tag{8}$$

Where $B(x, y)$ denotes the BLEU score of x and y; $l(x)$ denotes the length of x; $\text{F}_1(x, y)$ denotes F_1 score of x and y; $d_{\theta_j}(x)$ denotes the decode sequence of parameter θ_j which is generated from j-th MC iteration.

4. Experiments

We will refer to the four uncertainty metrics introduced in Section 3, i.e. deterministic sequence probability, MC-Dropout sequence probability, MC-Dropout sequence probability variance and MC-Dropout BLEU score variance as DSP, MCSP, MC-VAR and MC-BLEU-VAR respectively.

4.1. Comparison of Uncertainty Metrics

The settings of this experiment are that we have a pre-trained language model that has been fine-tuned on English QA using SQuAD v1.1 that we denote as M, and we also have a test set E (English, or *en*) that is expected to be in-distribution and a test set L expected to be out-of-distribution, each of size N. We then evaluate the uncertainties of these test sets (E and L) under M using different metrics ρ (DSP, MCSP, MC-VAR and MC-BLEU-VAR as described in Section 3). With the resulting lists containing ρ scores of each instance in E and L, we finally carry out two types of analysis for the uncertainty results:

1. Variation of F1 score on the test set as we gradually remove the most uncertain instances from the test set according to each ρ.
2. Distribution of each ρ on E and L.

Specifically, we used two languages for L: German (*de*) and simplified Chinese (*zh*), the reason being that *de* is closer to *en* while *zh* is far more distinct so one might expect a good out-of-distribution detector to yield slightly different uncertainty results for *de* and *zh*, although both should be detected as out-of-distribution. We chose each test set to have a size $N \approx 5000$, simply because the MLQA dataset contains in total 5029 QA instances in *de*, even though there are more for *en* and *zh*. For *en* and *zh*, we randomly sampled about 5000 QA instances from MLQA dataset.

In terms of models, we experimented with M in each of BERT-base (BERT), multi-BERT-base (m-BERT) (Devlin et al., 2018), and ALBERT (Lan et al., 2019). The motivation for using these models is that while m-BERT has seen *de* and *zh* in its pre-training stage, the other two models have never seen those two languages. Therefore, the degree to which *de* and *zh* are out-of-distribution should be similar for BERT and ALBERT, but they should appear less out-of-distribution for m-BERT.

4.2. Analysis of the Best Uncertainty Metric

As the difference between in and out-of-distribution instances is whether they are similar to the training data, we investigate what factors other than actually being inside training data contribute to such "similarity". We first analyse how our uncertainty metrics perform when language becomes the only difference in different test sets, i.e. when all other factors, including context, are kept the same.

Furthermore, we fix the language of the test set to *en*, and investigate how other factors affect the uncertainty of instances. Based on the result of the previous experiment we find MC-BLEU-VAR performs the best in detecting out-of-distribution data, we therefore extract QA instances with their MC-BLEU-VAR and take a deeper look at them.

We choose the results on BERT model because it showed the best results for detecting out-of-distribution instances. One hypothesis we make is that we assume that this metric represents information about question type, answer length

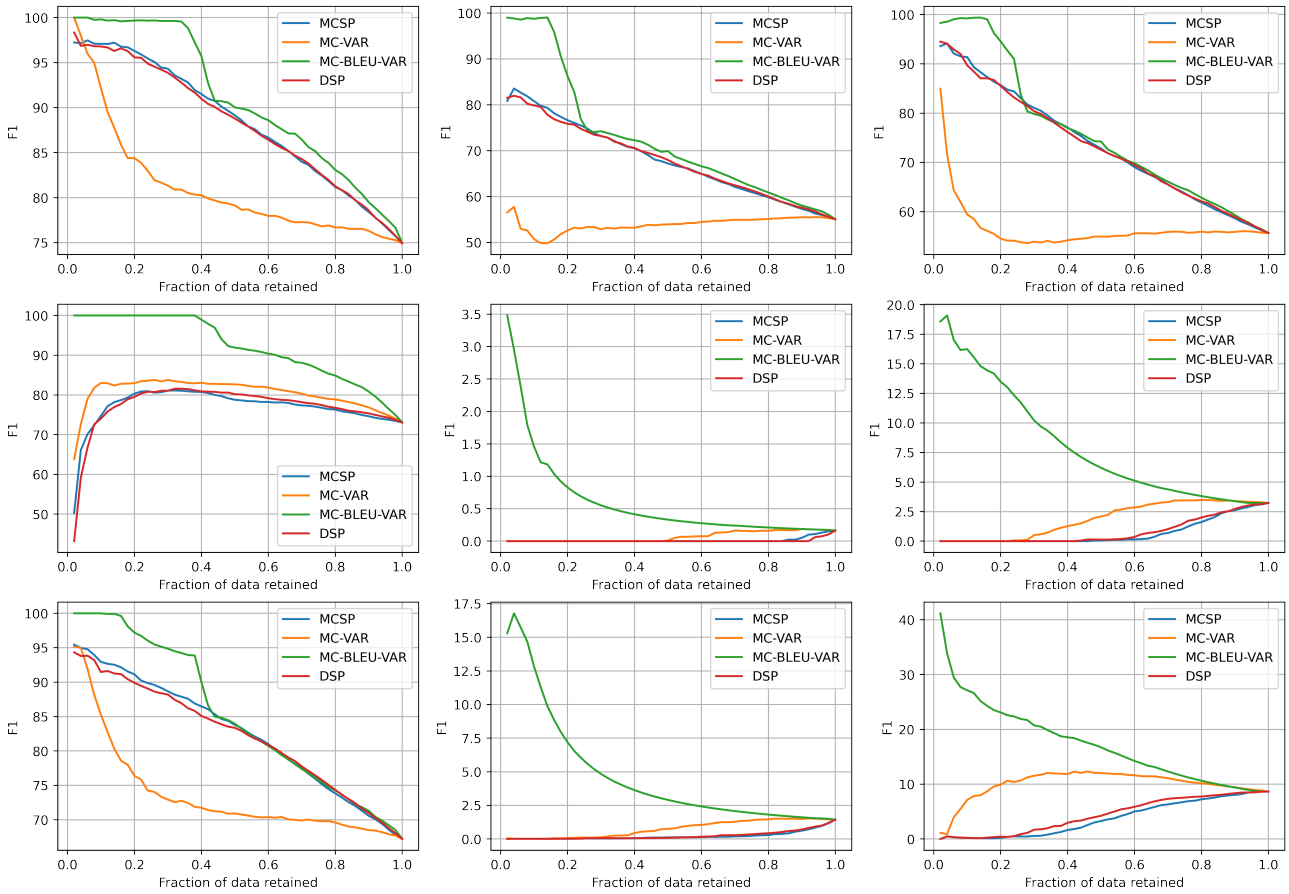

Figure 1. Comparisons of uncertainty metrics evaluated on in-distribution *en* (left), and out-of-distribution *de* (middle), *zh* (right) test sets using m-BERT (top 3), BERT (middle 3), ALBERT (bottom 3). X-axis is (1-fraction of data removed). The detailed discussion shows in the appendix

and context length, so we generate three features for each data: answer length, context length, question type (What, Which, Who, When, How, Other). For answer length and context length, we calculate their Spearmans correlation statistics with MC-BLEU-VAR score, and we use Analysis of Variance (ANOVA) table to show the relationship between question type and MC-BLEU-VAR score. We also check the corresponding answer type for each instance and do analysis.

4.3. Active Learning

In this experiment we will investigate the performance of all the uncertainty metrics in active learning. The MLQA dataset is fully labelled, so we can regard the ground truth answer as if were the data labeled by the human labelers. Using uncertainty as sample selection criteria in active learning, we need to decide whether the model is uncertain about a prediction: if it is, we will label it and feed it back to the model as training data. Specifically in this experiment, we will take the the most uncertain 15% data, and compare with the most certain 15% and random 15% data of the MLQA dataset that is sampled in Section 4.1 for each uncertainty metric from different models and languages as training data, using the training data excluding all re-training data as the test set. As such, the test data is the same for all the cases.

It can be seen that DSP, MCSP and MC-BLEU-VAR improve the F1 score most when the models are re-trained with their highest uncertainty data. The detailed active learning results and experiment discussion show in the appendix.

5. Conclusion

In this work we have compared four different uncertainty metrics for detecting out-of-distribution data in the task of multi-lingual question answering, and our results show that MC-BLEU-VAR outperforms the others. We further analysed the factors that contribute to the difference between in and out-of-distribution data, and found that language is the key factor. We then proposed a data selection strategy for active learning based on our developed uncertainty measure,

and compared it with another well-developed as well as a random strategy. The results showed that our method was more effective when selecting out-of-distribution data.

References

Abdi, A., Idris, N., and Ahmad, Z. Qapd: an ontology-based question answering system in the physics domain. *Soft Computing*, 22:213–230, 2018. doi: 10.1007/s00500-016-2328-2. URL https://doi.org/10.1007/s00500-016-2328-2.

Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 192—-199, 2000. URL http://www.faqs.org.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018. URL http://arxiv.org/abs/1810.04805.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, 2016. ISBN 9781510829008.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In *arXiv preprint arXiv:1909.11942*, 2019. URL https://github.com/google-research/ALBERT.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., and Schwenk, H. Mlqa: Evaluating cross-lingual extractive question answering. In *arXiv preprint arXiv:1910.07475*, 2019. URL http://arxiv.org/abs/1910.07475.

Liu, M., Tu, Z., Wang, Z., and Xu, X. Ltp: A new active learning strategy for bert-crf based named entity recognition. In *arXiv preprint arXiv:2001.02524v1*, 2020.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, 2002.

Peshterliev, S., Kearney, J., Jagannatha, A., Kiss, I., and Matsoukas, S. Active learning for new domains in natural language understanding. In *arXiv preprint arXiv:1810.03450*, 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2383–2392, jun 2016. URL http://arxiv.org/abs/1606.05250.

Srivastava, N., Hinton, G., Krizhevsky, A., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009, 2017. ISSN 10495258.

Xiao, T. Z., Gomez, A. N., and Gal, Y. Wat zei je? detecting out-of-distribution translations with variational transformers. *arXiv:2006.08344*, 2020.

Yan, Y., Nie, F., Li, W., Gao, C., Yang, Y., and Xu, D. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, 18 (12):2494–2502, 2016.

Zhang, Z., Yang, J., and Zhao, H. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*, 2020.

Appendix

A. Investigation Approach

In order to investigate and compare the ability of the four uncertainty measures to detect out-of-distribution data, we firstly evaluate the uncertainties of all instances of a QA test set according to all four metrics using a pre-trained transformer-based model that is fine-tuned on the task of QA in English with the SQuAD v1.1 (Rajpurkar et al., 2016) dataset. Once we have the evaluation outcomes, we then assess the quality of uncertainties given by each metric by removing their corresponding most-uncertain instances from the test set. In principle, if we remove the instances that our model is indeed uncertain about, the overall performance of the model on the remaining test set should improve (Xiao et al., 2020). Therefore, if the uncertainty measurement given by a metric is sensible, we expect an overall increase i n test-set F1 score as we remove more uncertain instances according to that metric. In addition, a good uncertainty measure should be able to detect test instances from a distribution that is novel to the model, so the distribution of uncertainty scores of in and out-of-distribution test data given by such metrics should be distinct.

We then further assess the metric that is found to be a good indicator of uncertainty to get some intuition of what the uncertainties are related to, and we do this by comparing certain statistics of the high-uncertainty and low-uncertainty instances. Finally, we explore the feasibility of using our best uncertainty metric for active learning, with which we expect to reduce the amount of annotated data needed for training to achieve certain performance.

B. Dataset

B.1. SQuAD v1.1

In order to prepare models to evaluate our uncertainty metrics, we used the Stanford Question Answering Dataset v1.1 (SQuAD v1.1(Rajpurkar et al., 2016)) to fine-tune the pre-trained language models on the question answering task in English. The reason for using this dataset for fine-tuning is that we intend to replicate the models trained in Lewis et al. (2019), where the authors also used SQuAD v1.1 to fine-tune pre-trained language models. Training on SQuAD v1.1 and evaluating our model on the MLQA dataset enabled us to use Lewis et al. (2019) as a reference to validate our training work, as shown in Table 1.

B.2. MLQA dataset

Presented recently in (Lewis et al., 2019), this dataset consists of substantial amount of QA instances mostly parallel in 7 languages. This parallel feature of this dataset is of particular interest to our work because it enables us to easily compare the uncertainties of the same QA instances in different languages thus facilitating the analysis of our uncertainty metrics.

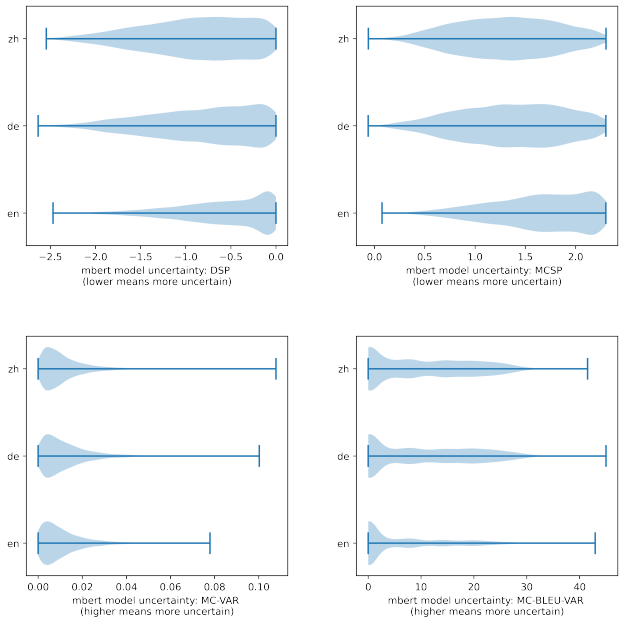

Figure 2. distribution of uncertainties for 3 languages

C. Results and Discussion

C.1. Comparison of Uncertainty Metrics

C.1.1. M-BERT

Fig. 2 shows the distribution of the four uncertainty metrics evaluated using m-BERT on the 3 languages. It can be seen that except MC-VAR, all metrics show different distributions on in-distribution (*en*) and out-of-distribution (*de*,*zh*) test sets. Specifically, for sequence probabilities (top 2 plots), most of *en* instances concentrate at far right while *de* and *zh* center at a lower sp values (indicating higher uncertainty). For MC-BLEU-VAR, although all 3 languages concentrated at 0, a lot of *de* and *zh* results spread to higher values (meaning higher uncertainty). Therefore, to some extent, all uncertainty measurements except MC-VAR managed to identify out-of-distribution data. However, such identification was not obvious, and this is expected because the m-BERT model "learned" *de* and *zh* in its pre-training stage, so they should not appear as totally out-of-distribution for m-BERT.

The top 3 plots in Fig. 1 shows how the performance of our m-BERT varies as we remove certain percentage of uncertain test data according to the four metrics. It should be noted that the very left part of the plots are of less interests because the test set is too small there so the resulting F1

score can be noisy. It is clear that if we remove uncertain data according to all of the metrics except MC-VAR, the performance is improved, indicating that those metrics did capture uncertainties as expected. It is also noted in these plots that the zero-shot performance of m-BERT on German and Chinese QA was not much worse than English, and we reckon this the beneficial outcome of multi-lingual pre-training of m-BERT, enabling it to grasp knowledge of many languages.

C.1.2. BERT

For a model that has never seen *de* and *zh*, MC-BLEU-VAR was able to detect these out-of-distribution test instances very well, as shown in bottom right plot of Fig. 4 . Regarding MC-VAR, although there are more uncertain instances of *de* and *zh* (indicated by farther tailing to the right) than *en*, the separation was not as sharp as MC-BLEU-VAR. On the other hand, the sp (both DSP and MCSP) metrics failed to identify in and out-of-distribution instances, with *en* having higher uncertainty than the other two. In addition, looking at the distribution of MC-BLEU-VAR (bottom right), it can be seen that the model is slightly less uncertain about *de* than about *zh* and this is what we expected because *de* is a lot closer to *en* (on which BERT was pre-trained) than *zh* is.

The middle row of Fig. 1 shows consistent results, with the F1 score constantly improving when we remove uncertain instances according to MC-BLEU-VAR. It is also noted that while the performance of BERT model on never-seen languages is generally poor, it is slightly better for *de* than for *zh*, and again this might be caused by the similarity between *en* and *de*.

C.1.3. ALBERT

As our ALBERT model, similarly to BERT, has never seen *de* and *zh*, these two languages should appear totally out-of-distribution to it. Therefore, we should expect similar results to those of BERT. Fig. 3 appear as expected, with MC-BLEU-VAR clearly separating the in and out-of-distribution instances and both sp metrics failing to assign correct uncertainties. The bottom row of Fig. 1 is consistent with our analysis, with the performance of the ALBERT model clearly improving as we take out the most uncertain test instances according to MC-BLEU-VAR. Again, in terms of both F1 score and out-of-distribution detection results, the difference between *de* and *zh* is consistent with our expectation.

C.2. Analysis of the Best Uncertainty Metric

In this part we investigate when the context is kept the same, i.e. when language is the only difference between test instances, how MC-BLEU-VAR performs. To achieve this we benefit from the parallel property of MLQA dataset, and

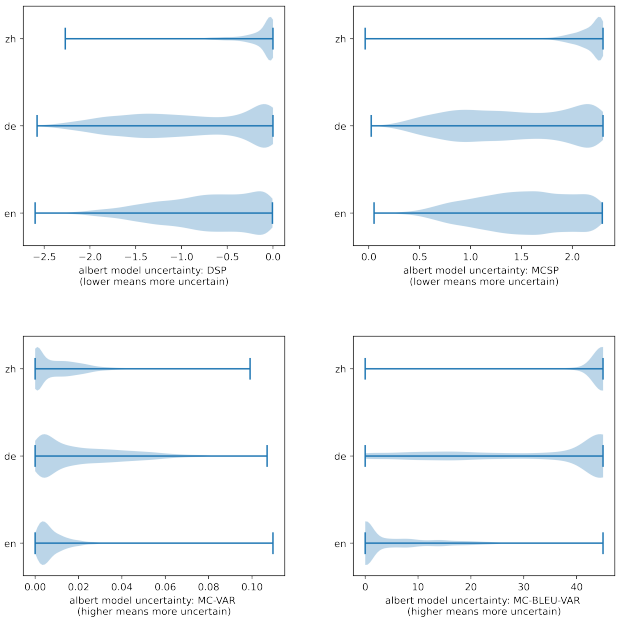

Figure 3. distribution of uncertainties for 3 languages

Table 2. Average MC-BLEU-VAR on each test set for 3 models. In the parenthesis is the difference between average MC-BLEU-VAR for *en* and for current language.

	en	de	zh
m-BERT	2.19	2.39 (0.20)	10.00 (7.81)
BERT	7.83	42.03 (34.20)	44.72 (36.89)
ALBERT	5.51	42.44 (36.93)	43.79 (38.28)

form 3 parallel test sets in *en*, *de* and *zh* respectively, each of size 1000. That is to say, every single QA instance in one of the three sets has an equivalent in each of the other two, with the same context but in different language. We record the average MC-BLEU-VAR score evaluated with each model on each test set in Table 2.

These results appear generally consistent with our analysis in Section 4.1. Specifically, for m-BERT the MC-BLEU-VAR score managed to identify the *zh* test set as out-of-distribution, but did not do so well for *de* which is much closer to *en*. Regarding the other two models who have never seen *de* and *zh*, MC-BLEU-VAR successfully detected the out-of-distribution instances when language is the only variant. Therefore, adding to our previous results, we can further conclude that the language is the key factor in terms of "similarity", and from the numerical result we confirm our prior belief the difference between average MC-BLEU-VAR *de* and *en* is smaller, as *de* and *en* are closer in terms of language family.

Table 3 shows the correlations of answer length and context

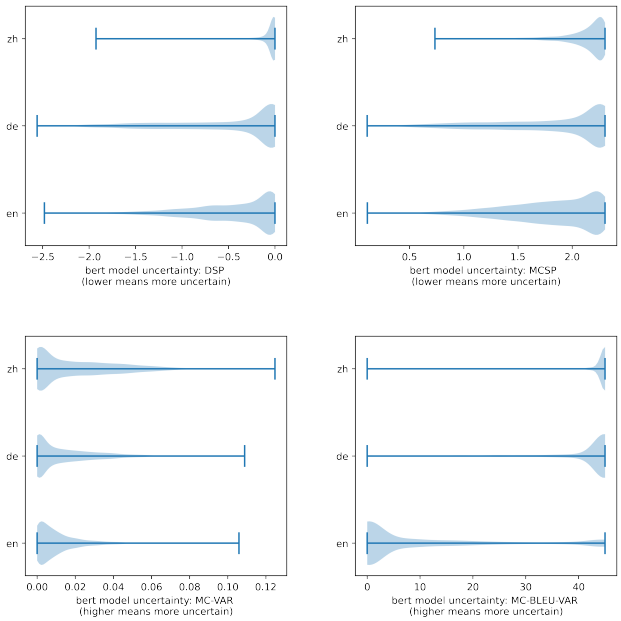

Figure 4. distribution of uncertainties for 3 languages

Table 3. Spearmans correlation between answer length, context length and MC-BLEU-VAR for *en* language on m-BERT model.

correlation	answer length	context length
m-BERT	0.164	0.144

length with MC-BLEU-VAR score. As is shown in the table, the statistics 0.164 and 0.144 indicate that both answer length and context length have weak positive dependent relationship with MC-BLEU-VAR, which means as length moves, either up or down, the score tends to move in the same direction. The ANOVA table for question type feature is shown in Table 5, the p-value 2.62E-08 indicate that MC-BLEU-VAR are dependent on question types, as we have enough evidence to reject the hypothesis that question type is irrelevant with uncertainty. To further check what kind of question has small score in general, we summarize each question type with mean, standard deviation in Table 4, the average score for the "when" question type is significantly smaller than the other types while the mean score for the "other" question type is much larger than the others' scores.

From Table 3 it is clear that QA instances with shorter answers tend to be assigned lower MC-BLEU-VAR scores, i.e. lower uncertainty. Checking those short-answer-length data in the low-score class (corresponding to lower uncertainty), we find almost all questions are asking for an entity name, a date, a number or a location. In contrast, the questions of the high score often ask for an explanation or a description of something, which requires a better context

Table 4. summary of different question types. N stands for total count of each question type, Mean and SD each stands for the score of mean and standard deviation.

Q Type	N	Mean	SD
how	640	8.28	13.50
other	410	11.22	15.62
what	2772	10.13	14.93
when	414	5.94	12.19
which	309	11.13	15.46
who	480	9.25	14.03

Table 5. ANOVA table for question type and MC-BLEU-VAR on m-BERT model.

	Sum_sq	df	F	Pr(>F)
Q Type	9.30E+03	5	8.78	2.62E-08
Residual	1.06E+06	5019	-	-

comprehension. So we could say this uncertainty metric is to some extent implicitly related to question difficulty.

C.3. Active Learning

C.3.1. Comparison among models

Table 6 shows the F1 scores of the three models re-trained with data selected using three different strategies. For the "H" and "L" methods Specifically, all four uncertainty metrics are concerned.

It can be seen that DSP, MCSP and MC-BLEU-VAR improve the F1 score most when the models are re-trained with their highest uncertainty data. MC-BLEU-VAR has the largest gap between the F1 score obtained by retraining the highest uncertainty data and the lowest ones, which can be seen from the "D" column in Table 6, suggesting that it is a better selection criterion than the other two. Therefore, in general, our proposed data selection strategy that is based on uncertainty obtained by different methods are satisfactory, shown by the fact that re-training the models with the most informative data (i.e. the most uncertain ones) increases the F1 score most significantly. However, it should be noted that since BERT and ALBERT are pre-trained only with *en*, the limited vocabulary makes the effect of active learning in terms of F1 score insignificant. In general, all the models perform as expected in most of the cases. The models re-trained with higher uncertainty data result in higher F1 scores. To some degree all the metrics are effective in terms of active learning.

Table 6. F1 scores of 3 models re-trained with data selected by 3 different selection strategies: H (15% with highest uncertainty), L (15% with lowest uncertainty), R (15% that is randomly sampled); D is the difference between the highest score and lowest score for each strategy. Mark the biggest difference

		DSP				MC-BLEU-VAR			
		H	L	R	D	H	L	R	D
ALBERT	en	74.53	63.00	68.89	11.53	74.35	59.99	69.81	***14.36***
	de	27.99	29.54	28.86	-1.55	27.97	22.9	29.13	***5.07***
	zh	9.64	11.25	10.26	-1.61	10.76	9.73	10.65	***1.03***
m-BERT	en	82.44	70.77	77.73	***11.67***	81.40	71.27	77.30	10.13
	de	65.57	54.24	61.98	11.33	65.18	51.03	61.86	***14.15***
	zh	65.29	55.09	60.80	10.20	64.03	50.23	60.53	***13.80***
BERT	en	81.67	77.62	78.43	4.05	82.93	72.53	78.78	***10.40***
	de	26.02	31.66	32.76	-5.64	27.77	24.24	33.42	***3.53***
	zh	10.95	12.05	12.52	-1.10	12.62	11.33	12.41	***1.29***
		MCSP				MC-VAR			
		H	L	R	D	H	L	R	D
ALBERT	en	74.20	63.09	70.28	11.11	71.37	66.74	69.71	4.63
	de	25.62	30.25	29.34	-4.63	29.95	29.36	28.00	0.59
	zh	9.45	11.26	10.53	-1.81	9.86	11.13	10.15	-1.27
m-BERT	en	82.04	70.77	77.23	11.27	78.24	74.85	76.96	3.39
	de	64.97	53.34	61.36	11.63	60.80	61.27	61.50	-0.47
	zh	63.76	55.04	60.27	8.72	60.53	62.50	60.50	-1.97
BERT	en	82.09	76.70	78.01	5.39	81.25	76.21	78.52	5.04
	de	24.99	30.83	34.07	-5.84	30.94	31.32	33.36	-0.38
	zh	11.20	12.61	11.79	-1.41	11.46	12.97	12.28	-1.51

C.3.2. Comparison with other strategies

Peshterliev (Peshterliev et al., 2018) proposed an active learning baseline strategy based on the score which corresponds to the model, and here in this experiment we will simply use the F1 score to choose the data for retraining, and compare such strategy with our proposed approach (i.e. selection based on MC-BLEU-VAR).

Table 7. F1 scores of m-BERT after training with strategy based on MC-BLEU-VAR and f1 score

	MC-BLEU-VAR	F1 score
en	82.00	83.12
de	65.22	56.99
zh	64.96	55.03

Table 7 shows the F1 score of m-BERT re-trained with the data selected by F1 score and by MC-BLEU-VAR. Although F1 score-based strategy has a slightly better performance in *en*, it has a worse performance in *de* and *zh* compared to the MC-BLEU-VAR strategy. This is because data with lower F1 scores are not necessarily the most informative ones. However, since test data in *de* and *zh* are out-of-distribution and the MC-BLEU-VAR strategy better measures such uncertainty, our strategy outperforms the F1 score strategy in above two languages. This also aligns with the conclusion of previous Appendix C.1.