Uncertainty as a tool to boost Knowledge based question answering

Anonymous Authors[1]

Abstract

Question Answering (QA) is one of the most widespread task in Natural Language Processing (NLP). Knowledge Base (KB) is a data structure that helps to extract answers more accurately and store facts efficiently.

This work focuses on the application of Uncertainty Estimation (UE) to extend the capabilities of the KBQA system. Using uncertainty estimation techniques, it is possible to sort answers by confidence and filter those questions where the model is less confident. It is demonstrated that uncertainty can enhance the performance of end-to-end KBQA system and a particular module of such system – Entity Linker.

Experiments included various uncertainty estimation approaches based on single model and ensemble estimations on different QA datasets. It is also shown that uncertainty estimates can be used to study the behavior of the model when answering different types of questions.

1. Introduction

Question Answering (QA) is a very hot and complex topic in Natural Language Processing (NLP). There are different approaches to retrieve entities as answers for the questions from Knowledge Base (KB) of Knowledge Graphs (KG). Such structures add efficiency in storing information (Huang et al., 2019) and can improve result especially in case of simple questions (Yani & Krisnadhi, 2021). Everyone faces with such systems while making queries in the browser which indicates not only the scientific but also the industrial relevance of the topic. Moreover, implementations of QA models are quite in demand, which is evidenced by the fact that they are presented on Hugging Face.

Uncertainty Estimation (UE) technics are widely used in medicine (Ayhan et al., 2020), autonomous driving (Loquercio et al., 2020), NLP (Xiao & Wang, 2019), Computer Vision (CV) (Kendall & Gal, 2017), etc. UE methods allow more accurate estimations and filtering answers, provide user with a score of model's uncertainty. This may be very important in many cases, especially when the cost of mistake is high.

The problem is that even modern KBQA systems are not perfect, while there are potential ways to improve their work, one of which is UE. Although there is a prospect of applying uncertainty estimation methods to the topic of Question Answering over Knowledge Base, there is rather little research on this topic, which opens research possibilities. But even existing research show us the potential of this field (Zhang et al., 2021) where authors introduce bayesian end-to-end model with estimated uncertainties for simple question answering over knowledge bases or more broadly applictaion of uncertainty estimation methods for Natural Language Processing, especially using large Transformer models (Shelmanov et al., 2021).

2. Related work

Uncertainty Estimation and Knowledge Based Question Answering are well developed separately, and then we will briefly review the literature on these tow topics.

2.1. KBQA

In recent years knowledge bases grows to such sufficient volumes that might be used as a very useful source of information. Graph is a data structure that helps efficiently store information. It is preferable to use structured information in many cases, thus knowledge graphs become a great option to build a knowledge base. It is especially appropriate for QA task because knowledge graphs consist of data represented in Resource Description Framewor (RDF) format. The idea is that RDF data is a set of triplets: (subject, property, object). In QA case we can show it on example:

The fact - "Jimmy Hendrics played on guitar" will have the following triplet: (subject = Jimmy Hendrics, property = play on instrument, object = guitar)

But it is worth mention that in such representation using words there are a lot of ambiguity issues that's why each

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. **AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

element in triplet is represented using unique ID.

The most well-known knowledge bases built on knowledge graphs are Wikidata (Vrandečić & Krötzsch, 2014) which contains already more than 70 million entities with links and DBPedia (Lehmann et al., 2015) that possess more than 39 millions of links.

KBQA approaches

First of all, we list the main approaches that exist in the KBQA problem: Sematic Parsing (SP) and Information-Retrieval (IR), end-to-end neural approach which has only recently become widespread mainly after transformer architecture appeared and distance-based methods.

Historically SP appeared firstly. It's idea basing on the question converting into query to Knowledge Base 1. Example of application of such methods demonstrated in (Bao et al., 2014) and considered under Machine-Translation scope. SP requires a specific markup of the training data that is computationally expensive. Besides, SP does not work well on out-of-domain (OOD) questions, which is a big problem for question answering model.

IR approach is based on NN with some manual rules 1. One of examples of such approach is demonstrated in (Yin et al., 2017).

As discussed if (Lan et al., 2022) for complex questions IR and SP is more popular approaches.

End-to-end approach has advantage that in fact one model do all the stuff. Let us list some examples of such scientific articles ((Lukovnikov et al., 2017); (Zhou et al., 2021)).

As mentioned before density based approach have also become widespread within the framework of the KBQA task. One of the most famous articles within this topic is (Huang et al., 2019).

In recent years some synthetic approaches also appears, for example application of transformer model with distance based method to KBQA task illustrated in (Baramiia et al.). In this article authors propose an approach to rank answers using deep NN transformer model over knowledge graph basing on distance between answers in form of embeddings.

Density based methods in comparison with SP have an advantage that it performs well on OOD questions.

Also, since knowledge graph may have very complex structure it is worth mention that representation learning is actively used to preprocess original graph in form of graph embeddings.

KBQA topic is discussed in many articles, a survey among them is presented in (Lan et al., 2022). There are studies that demonstrate that there is a potential gain of adding UE methods inside KBQA system ((Cui et al., 2019), (Zhang et al., 2021)).

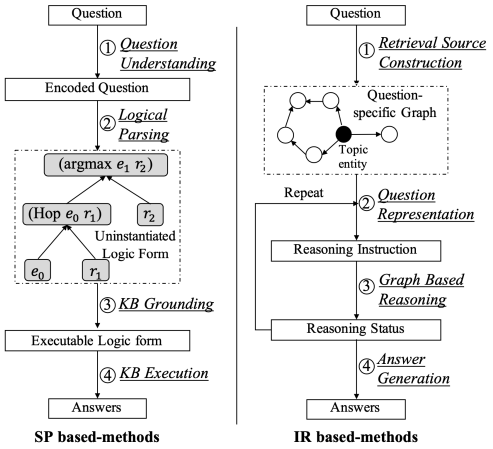

Figure 1. Illustration of IR and SP approaches. Source: (Lan et al., 2022)

2.2. Entity Linking

Quite often QA system consist of different modules (however T5 model we discuss for Question answering is end-to-end). One of the basic modules is Entity Linker (EL). This module extracts key components from the text like subject and link it with information from data base correctly. The subject is used as a part of a request to the knowledge base, in response to which one can get an answer, thus creating a QA system.

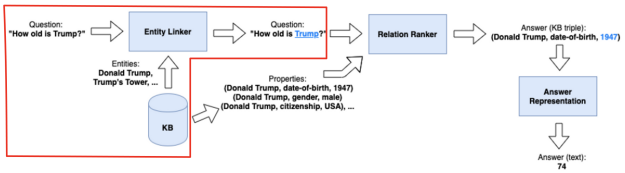

Figure 2. Illustration of Entity linking task on simple example

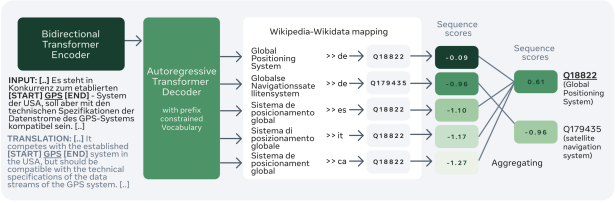

Figure 3. mGENRE architecture. Source of illustration is original mGENRE article: (De Cao et al., 2022)

2.3. Uncertainty Estimation

Basic idea for the uncertainty estimation is to provide some value for a particular input that indicates the model's confidence or uncertainty in answer. Uncertainty could be

estimated in several ways. One of them is Monte Carlo (MC) dropout technique (Gal & Ghahramani, 2016) that connects dropout with Bayesian theory and ensembles. Another approach is to construct deep ensembles described in (Lakshminarayanan et al., 2017) which often demonstrate high quality of uncertainty measurement.

It is worth to mention here that there exist different types of uncertainty: data-based (also known as aleatoric or irreducible) and model-based (also known as epistemic or reducible) uncertainty. Such differentiation is useful because it helps better understand nature of total uncertainty via analysing its parts: aleatoric and epistemic. Below we will illustrate that there are particular methods of uncertainty estimate that are sensitive to particular type of uncertainty.

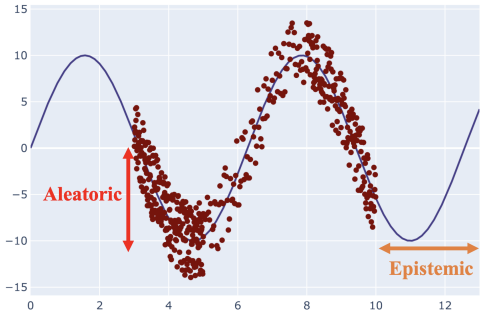

Figure 4. Illustrtion the difference between aleatoric and epistemic uncertainty. Source of image (Abdar et al., 2021)

3. Dataset Description

One of the most popular datasets for KBQA task is Simple Questions probosed by (Bordes et al., 2015). It is first dataset that contained information about knowledge graph's triplets. MKQA is another core KBQA multilingual dataset (26 languages) that has Wikidata unique indexes made by Apple (Longpre et al., 2021). Such links gave an opportunity to consider KGQA task on this dataset. For our experiments we additionally took subsample from SQ dataset which questions possess links to current wikidata.

For our experiments we are going to consider RuBQ 2.0 dataset (Rybin et al., 2021). This is an extention of RuBQ dataset (Korablinov & Braslavski, 2020) and has 2 languages: English and Russian. Russian variant is achieved via machine translation from English. For each question there is a corresponding SPARQL query with wikidata ID. Also, in order to check the performance of the model, we will conduct experiments on the well-established in this topic, as mentioned above, the Simple Questions dataset. We will do experiments with RuBQ 2.0 with only 1-hop (direct) questions and with the whole test part with annotated questions to wikidata id.

Another dataset, that attracts out attention is Mewsli-9, since this dataset appeared in original article for mGENRE entity linker on which SOTA multilingual entity linking results was obtained by mGENRE.

4. ML Methods and algorithms

4.1. Uncertiainty metrics

As UE measures, we are going to use: We will start with **Entropy-based measures**:

- **Entropy**

  Entropy is one of the most widespread measures of uncertainty. The higher the entropy value, the higher the uncertainty. Basically, it is used to calculate uncertainty of a single model.

$$Entropy = -\sum_{i=1}^{N} p_i log_2(p_i)$$

  Where N is total number of classes and p_i – is probability of belonging to class i. But it is possible to calculate ensemble version of Entropy by arevaging entropies of each model.

- **Predictive Entropy**

  Predictive Entropy is an ensemble-based measure of uncertainty that estimates a predictive distribution obtained within ensemble's inference.

$$PEntropy = -\sum_{i=1}^{N} \left(\left(\sum_{k=1}^{K} p_{ik} \right) \cdot log_2 \left(\sum_{k=1}^{K} p_{ik} \right) \right)$$

  Thus, we average probabilities for each class between all models and compute classical entropy over obtained averaged probabilities. It is generally accepted that predictive entropy is a measure of overall model uncertainty, that is, it is a measure that is not sensitive to aleatoric uncertainty (uncertainty in the data) and is also not sensitive to epistemic uncertainty (model uncertainty).

- **Expected Entropy**

$$EEntropy = -\sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \cdot log_2 p_{ik}$$

  Here we compute K entropies and after that compute the sum over them. Expected entropy is a classical measure of data-based uncertainty

- **Bayesian Adversarial Training by Disagreement (BALD) or Mutual Information (MI)**

  This method was proposed by (Houlsby et al., 2011) and can be defined for our task as:

$$BALD = PEntropy - EEntropy$$

  As mentioned above since Predictive Entropy is a measure of total uncertainty and Expected Entropy is a measure of data-based uncertainty because total uncertainty consists of data and model-based uncertainty it is assumed that the mutual information (MI) between model parameters and input data, also called BALD, is a measure of model-based uncertainty.

Methods **based directly on probabilities**:

- **Score based estimation**

  Maximum probability is the most straightforward measure of uncertainty. This measure is reverse to the confidence of the model in the most probable class (which in fact top-1 probability)

$$MaxProb = -top_1(pred\_distr(q))$$

  Where pred_dist – (predictive distribution) is a set of probabilities obtained during forward pass from the model for question – q.

- **Delta – based estimation**

  Difference between top-1 and top-2 probabilities can be used as uncertainty measure, also known as Delta. The purpose of this metric is to estimate how far the mode of the predictive distribution differs from the nearest probability to it (second).

$$Delta = -top_1(pred\_distr(q)) + top_2(pred\_distr(q)) \quad (1)$$

  Where pred_dist – (predictive distribution) is a set of probabilities obtained during forward pass from the model for question – q.

Additional methods:

- **Expected Pairwise Kullback – Leibler divergence**

  The idea is to calculate KL divergence for each pair of predictive distributions for K models. Similar to BALD, EPKL shows how much the models in the ensemble diverge in their predictions in terms of classical approach to measure distance between distributions – KL divergence, that is, it is also a measure of model-based uncertainty.

$$EPKL = \binom{M}{2}^{-1} \sum_{i=1}^{M} \sum_{j=1}^{M} I(i \neq j) KL(P(\mathbf{y}|\mathbf{x}, \theta_i) || P(\mathbf{y}|\mathbf{x}, \theta_j))$$

  where M - number of models in ensemble, $P(\mathbf{y}|\mathbf{x}, \theta_i)$ - predictive distribution of $i-th$ model parameterized by θ_i.

Of course there are also other metrics of uncertainty that might be estimated, for example density based metrics: NUQ (Kotelevskii et al., 2022), DDU (Mukhoti et al., 2021). But we will not focus on density based metrics in this research. Here we analyse ensemble based and score/delta based metrics.

4.2. Quantification of the quality of Uncertainty Estimation via Area Under Rejection Curve.

Area Under Rejection Curve as an approach to compare quality of uncertainty estimation. This curve demonstrates what is the gain in quality we achieve when we filter our sample from the questions where our model is not confident.

Absolute Area Under Rejection Curve (AAURC) - is area under a rejection curve obtained by a particular uncertainty measure. The higher this value, the better. Example of such curve demonstrated on 5

Another way is to calculate only area that is higher than baseline quality on the whole set. The idea is to measure gain obtained only by filtering samples using uncertainty estimation. Let's call this area by Alternative Area Under Rejection Curve (AAURC), illustration is available on 6. This value might be useful in situation when we want to distinguish baseline quality of the model from surplus obtained by uncertainty estimation.

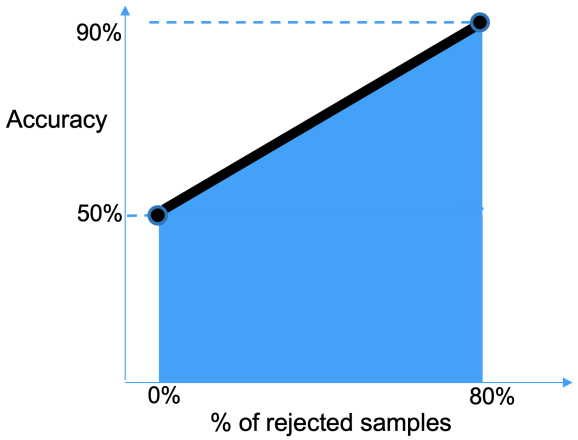

Figure 5. Illustration of Absolute Area Under Rejection Curve. For this example its value equals to $0.8 \cdot (0.9 + 0.5) \cdot 0.5 = 0.56$ (56%)

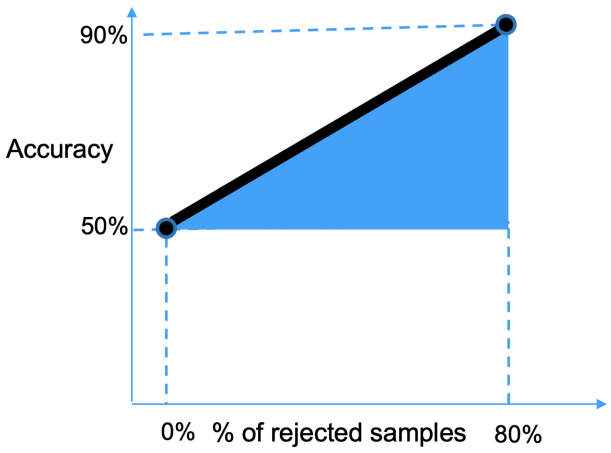

Figure 6. Illustration of Relative Area Under Rejection Curve. For this example its value equals to $0.8 \cdot (0.9 - 0.5) \cdot 0.5 = 0.16$ (16%)

4.3. Monte Carlo Dropout

Base theory

Dropout (Srivastava et al., 2014) is a classical regularization technique to prevent neural networks from overfitting, moreover (Gal & Ghahramani, 2016) introduce theoretical framework casting training process of deep neural networks with dropout as approximate Bayesian inference in deep Gaussian processes. As a result dropout could be turned on not only during training stage but also during inference to obtain the approximations of Bayesian inference.

More formally for $t - th$ layer output where t from $\{1, T\}$, $x_t = activation(x_{t-1}|W_t, M_t)$, where M_t is a mask - Bernoulli random variable with parameter $1 - p$. Final output is sampled via formula $f_k(x) = f(x|\{M_t^{(k)}\})$ as discussed in (Shelmanov et al., 2021), where k is number of passes through unique mask.

Although (Lakshminarayanan et al., 2017) claim that deep ensembles outperform MC-Dropout, MC-Dropout is much easier to obtain since it requires less computational resources because there is no need to train several independent models, that is why it is actively used in many articles (Gal & Ghahramani, 2016), (Shelmanov et al., 2021), etc.

MC-dropout ensemble application for seq-to-seq QA task

Mainly the main problem is to conduct proper experiments related to estimated uncertainty in sequence-to-sequence model within KBQA system.

On the one hand, we can consider the problem of answering a given question as a problem of classification with a huge number of possible classes - answers.

On the other hand, the seq-to-seq approach imposes difficulties for this approach due to the way the answer is generated. The point is that when we ensemble the model using the MC-dropout approach, when limiting the output to n passes, we may obtain class probabilities that are not present in other predictions, but nevertheless all classes must be considered in the uncertainty calculation. This is thus different from the classical approach to calculating uncertainty for the classification problem. In this paper, we will follow the concept of Expectation of Products (EoP) proposed in (Malinin & Gales, 2020).

Of course, since the number of classes if extremely huge it is an additional problem for us. Additionally, since research interest considers not only end-to-end question answering system but also such an important module of many IR QA systems – Entity Linker separately.

5. Experiments

5.1. Entity Linking

Summary of Entity Linking experiments are illustrated on Figure 7 and Figure 8. We can see that there is no obvious leader in terms of uncertainty measure, however distribution over languages demonstrates that it is much easier for model to cope with entity linking task for Russian, English and German languages.

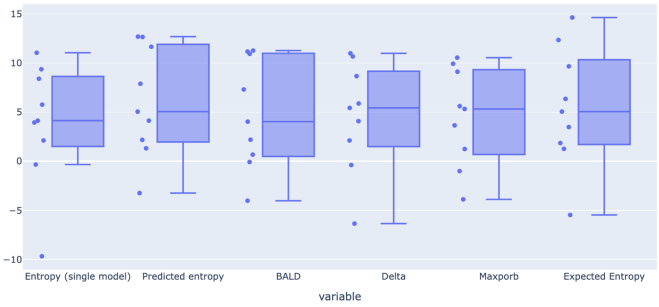

Figure 7. Distribution of Relative Areas Under Rejection Curves over languages for different UE measures (area over y-axis in %)

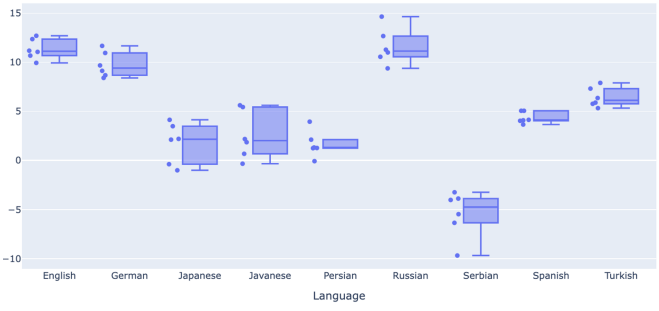

Figure 8. Distribution of Relative Areas Under Rejection Curves over UE measures over different measures of UE for languages (area over y-axis in %)

Results for EL part are the following:

- Experiments demonstrate that uncertainty quantification could be efficiently used for the task of entity linking in case of mGENRE model. This is especially obvious for English, German, Russian. Although, there are cases when uncertainty estimation does not help, for example for Serbian and in some cases for Javanese, Japanese and Persian.

- The dataset also matters, because the structure differs, for example model may perform better on RUBQ 2.0 and Simple Questions because they consist of short questions of widespread languages (English and Russian).

- There is no significant leader in terms of metric of uncertainty, but every time both for absolute area under rejection curve and area under curve added by uncertainty integration predicted entropy showed the best results among all metrics almost each time as for information received on aggregated AUC data.

- mGENRE performs well on the task of entity linking in multilingual case. But end-to-end question answering system could not be realized using only mGENRE, even if we help it with such strong NER as Stanza.

All the areas under curve for this task demonstrated on the Figure 10.

5.2. End-to-End KBQA based on T5 model

Our experiments were conducted basing on the t5-xl-ssm-nq model (Raffel et al., 2020) from Hugging Face. The advantage of models from this source is fast reproducibility. T5 that we used (T5-XL-SSM-NQ) was pre-trained using T5's denoising objective on C4, subsequently additionally pre-trained using REALM's salient span masking objective on Wikipedia, and finally fine-tuned on Natural Questions (Kwiatkowski et al., 2019).

All obtained absolute ares under curves for KBQA task presented in Table 3. Baseline top-1 accuracy for each dataset represented in Table 2 and Baseline top-n accuracy is demonstrated in Table 1. Now let's summarise results from these tables.

Summary of results on UE T5 part:

For Simple Questions we observe the following results:

- Expected entropy – worse behavior on high rejection rates

- Expected entropy – worse behavior on high rejection rates Predictive entropy, EPKL, BALD stable and very similar

- Best performance (especially for low rejection rates) for Delta, Maxprob and Eentropy. (All of them perform worse on high rejection rates)

For RuBQ 2.0 we observe the following results:

- Expected entropy – worse behavior on high rejection rates

- Predictive entropy, EPKL, BALD stable and very similar, but gain is small.

- Best performance by Delta, Maxprob

- Eentropy perform very well at low rejection rates

- Model fails on questions w/o answer

Also we observe the following total trends:

- Single model-based metrics often demonstrate good performance. They outperform ensemble metrics on RuBQ 2.0 dataset and SQ in graph part.

- Almost each time score is the best Single model-based metric from Single model-based metrics.

- Ensemdle-based models outperform single-based models only on SQ full test dataset.

- Gain from using trie during hypothesis generation is very negligable.

- Changing dropout rate reduces results.

6. Discussion

As we understood from this research uncertainty can help enhance KBQA system and its module - Entity Linker. There is evidence that simple straightforward metrics based on single model estimation could become a very strong baseline and ensemble metrics are not always better. There are some topics that might be developed based on uncertainty estimates: mixture of experts, snapshot ensembles, density based estimation.

7. Conclusion

Uncertainty estimation could be efficiently used to boost KBQA system performance and it's sufficient module - Entity Linker. Within this research we demonstrated this improvement for KBQA basing on T5 model and for EL basing on mGENRE. We have to keep in mind that there might be cases when uncertainty does not help. Another important point is the importance of data that we use. With complex

questions and statements uncertainty measures copes much worse. For different tasks there might be different leader in terms of metric of uncertainty, for example we obtained that there is no clear leader in terms of uncertainty metrics for EL task, but single based model more often and in particular score was the leader in end-to-end KBQA task.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Ayhan, M. S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., and Berens, P. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724, 2020.

Bao, J., Duan, N., Zhou, M., and Zhao, T. Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 967–976, 2014.

Baramiia, N., Rogulina, A., Petrakov, S., Kornilov, V., and Razzhigaev, A. Ranking approach to monolingual question answering over knowledge graphs.

Bordes, A., Usunier, N., Chopra, S., and Weston, J. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

Cui, W., Xiao, Y., Wang, H., Song, Y., Hwang, S.-w., and Wang, W. Kbqa: learning question answering over qa corpora and knowledge bases. *arXiv preprint arXiv:1903.02419*, 2019.

De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., and Petroni, F. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290, 2022.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Huang, X., Zhang, J., Li, D., and Li, P. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 105–113, 2019.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Korablinov, V. and Braslavski, P. Rubq: a russian dataset for question answering over wikidata. In *International Semantic Web Conference*, pp. 97–110. Springer, 2020.

Kotelevskii, N., Artemenkov, A., Fedyanin, K., Noskov, F., Fishkov, A., Petiushko, A., and Panov, M. Nuq: Nonparametric uncertainty quantification for deterministic neural networks. *arXiv preprint arXiv:2202.03101*, 2022.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., and Wen, J.-R. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

Longpre, S., Lu, Y., and Daiber, J. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021.

Loquercio, A., Segu, M., and Scaramuzza, D. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

Lukovnikov, D., Fischer, A., Lehmann, J., and Auer, S. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*, pp. 1211–1220, 2017.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. Deep deterministic uncertainty: A simple baseline. *arXiv e-prints*, pp. arXiv–2102, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Rybin, I., Korablinov, V., Efimov, P., and Braslavski, P. Rubq 2.0: an innovated russian question answering dataset. In *European Semantic Web Conference*, pp. 532–547. Springer, 2021.

Shelmanov, A., Tsymbalov, E., Puzyrev, D., Fedyanin, K., Panchenko, A., and Panov, M. How certain is your transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1833–1840, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57 (10):78–85, 2014.

Xiao, Y. and Wang, W. Y. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7322–7329, 2019.

Yani, M. and Krisnadhi, A. A. Challenges, techniques, and trends of simple knowledge graph question answering: a survey. *Information*, 12(7):271, 2021.

Yin, J., Zhao, W. X., and Li, X.-M. Type-aware question answering over knowledge base with attention-based tree-structured neural networks. *Journal of Computer Science and Technology*, 32(4):805–813, 2017.

Zhang, L., Lin, C., Zhou, D., He, Y., and Zhang, M. A bayesian end-to-end model with estimated uncertainties for simple question answering over knowledge bases. *Computer Speech & Language*, 66:101167, 2021.

Zhou, G., Xie, Z., Yu, Z., and Huang, J. X. Dfm: A parameter-shared deep fused model for knowledge base question answering. *Information Sciences*, 547:103–118, 2021.

A. Tables and Figures with results

KGQA Method	Beam Size	SQ Full Test	SQ In graph	RuBQ 1186	Full RuBQ Test 1809
Single T5-XL-SSM-NQ	10	18.86	18.95	51.35	31.62
	20	21.72	22.39	54.22	31.67
EoP Ensemble T5-XL-SSM-NQ	10	22.40	18.95	42.83	33.06
	20	26.06	22.19	45.28	35.16

Table 1. KGQA: Total accuracy (accurate if correct entity is predicted anywhere among all top-beam size candidates)

KGQA Method	Beam Size	SQ Full Test	SQ In graph	RuBQ 1186	Full RuBQ Test 1809
Single T5-XL-SSM-NQ	10	7.28	8.29	34.32	20.12
	20	7.41	8.92	34.40	20.18
EoP Ensemble T5-XL-SSM-NQ	10	9.39	8.29	27.32	21.12
	20	9.47	8.2	26.98	20.56

Table 2. KGQA: Top-1 accuracy (accurate if top-ranked candidate is the correct entity)

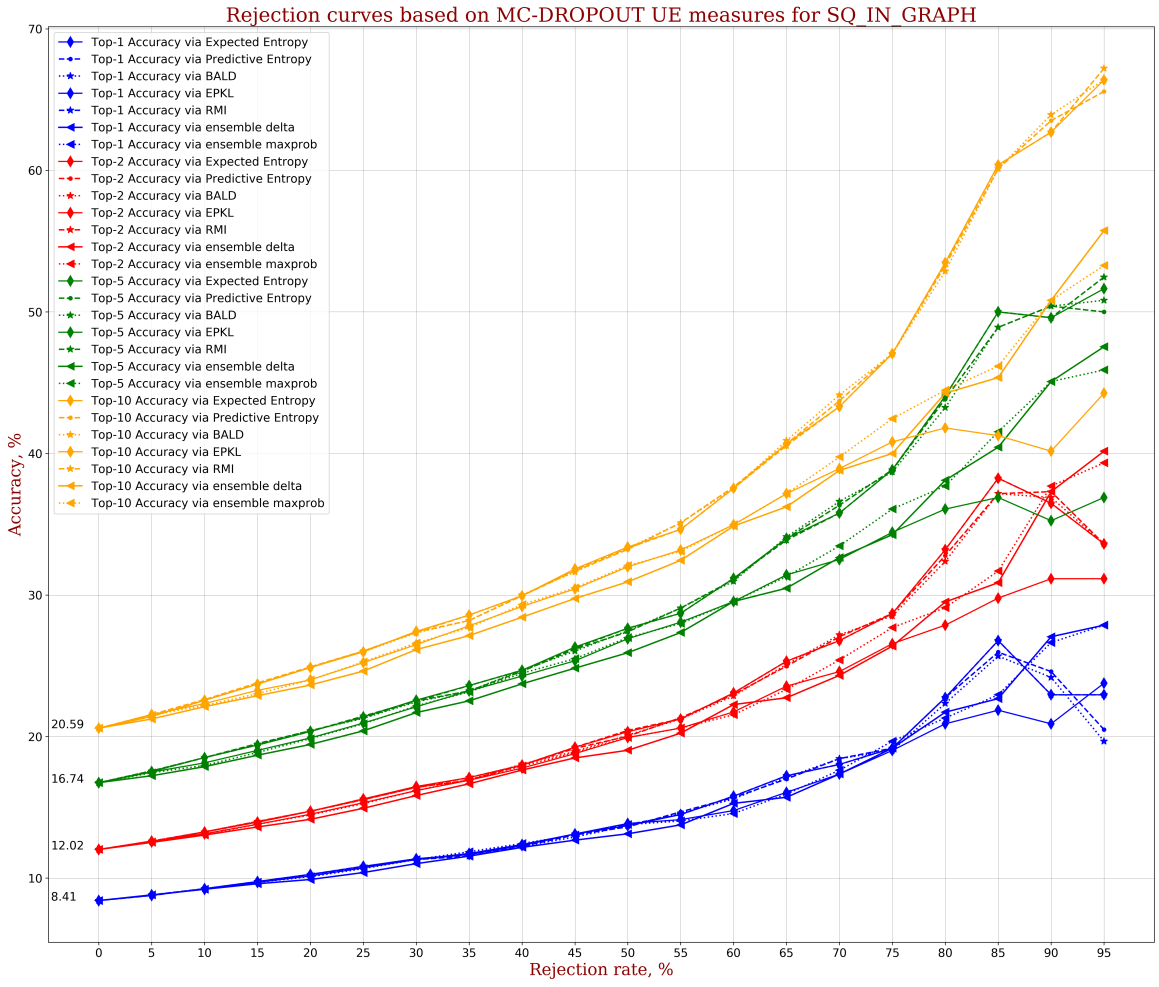

Figure 9. MC-Dropout rejection curve for SQ IN Graph dataset

					Absolute AUCs							Comparative AUCs						
Dataset	Language	EL/QA	With/without NER	Number of answers	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected Entropy	MEAN	Entropy (single model)	Predicted entropy	BALD	Delta	Maxprob	Expected entropy	MEAN
Mewsli 9	German	EL	with_Stanza	full	50.81	53.61	53.17	53.18	52.9	52.71	52.73	8.73	11.53	11.09	11.1	10.82	10.63	10.65
			wo_NER	full	44.65	48.36	47.36	42.8	43.99	45.27	45.4	8.1	11.81	10.81	6.25	7.44	8.72	8.86
	English	EL	with_Stanza	full	40.75	40.4	38.93	38.41	37.31	39.92	39.29	11.43	11.08	9.61	9.09	7.99	10.6	9.97
			wo_NER	full	41.93	38.42	36.95	40.14	38.41	37.92	38.96	12.61	9.1	7.63	10.82	9.09	8.6	9.64
	Spanish	EL	with_Stanza	full	28.49	31.12	30.71	30.45	30.04	31.24	30.34	1.71	4.34	3.93	3.67	3.26	4.46	3.56
			wo_NER	full	41.43	40.64	38.99	39.36	38.92	40.51	39.98	6.58	5.79	4.14	4.51	4.07	5.66	5.12
	Persian	EL	with_Stanza	full	46.25	44.28	44.35	43.73	43.44	44.17	44.37	3.75	1.78	1.85	1.23	0.94	1.67	1.87
			wo_NER	full	39.85	36.57	33.71	38.72	37.25	36.57	37.11	4.15	0.87	-1.99	3.02	1.55	0.87	1.41
	Japanese	EL	with_Stanza	full	33.83	35.08	33.6	30.46	30.24	33.78	32.83	2.38	3.63	2.15	-0.99	-1.21	2.33	1.38
			wo_NER	full	35.02	37.79	35.4	33.39	32.37	37.79	35.29	1.87	4.64	2.25	0.24	-0.78	4.64	2.14
	Serbian	EL	wo_NER	full	52.8	59.24	58.46	56.13	58.6	57.01	57.04	-9.68	-3.24	-4.02	-6.35	-3.88	-5.47	-5.44
	Javanese	EL	wo_NER	full	52.37	54.88	53.38	58.14	58.32	54.56	55.28	-0.33	2.18	0.68	5.44	5.62	1.86	2.58
	Turkish	EL	with_Stanza	full	43.28	45.21	44.47	43.75	41.98	45.66	44.06	7.16	9.09	8.35	7.63	5.86	9.54	7.94
			wo_NER	full	45.6	47.92	47.5	45.37	46.02	44.41	46.14	4.38	6.7	6.28	4.15	4.8	3.19	4.92
RUBQ 2.0	Russian	EL	wo_NER	single_answer	57.71	60.01	58.42	57.91	56.42	60	58.41	8.84	11.14	9.55	9.04	7.55	11.13	9.54
	English	EL	with_Stanza	single_answer	50.26	53.62	51.53	53.24	51.55	52.36	52.09	9.88	13.24	11.15	12.86	11.17	11.98	11.71
			wo_NER	single_answer	54.73	57.16	54.67	54.28	54.67	58.31	55.64	12.65	15.08	12.59	12.2	12.59	16.23	13.56
	Russian	EL	with_Stanza	single_answer	50.72	55	53.81	53.76	54.38	58.96	54.44	9.92	14.2	13.01	12.96	13.58	18.16	13.64
Simple Questions	English	EL	With_Stanza	full	61.54	68.09	67.15	65.14	64.06	11.35	56.22	14.36	20.91	19.97	17.96	16.88	20.6	18.45
				single_answer	59.1	63.13	61.37	57.62	56.85	62.38	60.08	17.02	21.05	19.29	15.54	14.77	20.3	18
			wo_NER	full	22.81	23.04	22.99	21.74	21.56	21.05	22.2	7.51	7.74	7.69	6.44	6.26	5.75	6.9
				single_answer	12.79	13.19	11.32	10.28	10.55	14.63	12.13	3.01	3.41	1.54	0.5	0.77	4.85	2.35
		QA	With_Stanza	full	4.4	3.67	3.63	3.1	3	3.74	3.59	-1.98	-2.71	-2.75	-3.28	-3.38	-2.64	-2.79
				single_answer	7.78	7.04	8.45	11.68	11.38	7.18	8.92	-7.52	-8.26	-6.85	-3.62	-3.92	-8.12	-6.38
			wo_NER	single_answer	9.02	10.69	11.2	10.21	11.32	7.36	9.97	-4.58	-2.91	-2.4	-3.39	-2.28	-6.24	-3.63
				full	4.67	3.21	2.66	4.53	3.71	4.21	3.83	-0.43	-1.89	-2.44	-0.57	-1.39	-0.89	-1.27

Figure 10. Total results for Entity Linfing

Model	Beam Size	UE	SQ Full Test	SQ In Graph	RuBQ 1186	RuBQ Full Test 1809
Single T5-XL-SSM-NQ	10	score	**0.1583**	**0.1510**	**0.5569**	**0.4648**
		delta	0.1545	0.1475	0.5559	0.4630
		entropy	0.1574	0.1492	0.5522	0.4597
	20	score	0.1606	0.1510	0.5584	0.4627
		delta	0.1577	0.1483	0.5583	0.4621
		entropy	0.1608	0.1500	0.5508	0.4560
EoP ensemble T5-XL-SSM-NQ	10	score	0.1477	0.1381	0.4445	0.3660
		delta	0.1434	0.1321	0.4579	0.3753
		eentropy	0.1325	0.1268	0.3830	0.3086
		pentropy	0.1673	0.1472	0.4168	0.3430
		bald	0.1683	0.1484	0.4143	0.3418
		epkl	0.1680	0.1480	0.4249	0.3517
		rmi	0.1682	0.1477	0.4249	0.3517
	20	score	0.1537	0.1394	0.4252	0.3306
		delta	0.1471	0.1349	0.4404	0.3493
		eentropy	0.1412	0.1276	0.3609	0.2808
		pentropy	0.1651	0.1426	0.4481	0.3031
		bald	0.1640	0.1415	0.4484	0.3017
		epkl	0.1655	0.1434	0.4485	0.3078
		rmi	0.1656	0.1436	0.4484	0.3090

Table 3. Absolute area under rejection curve (top-1 accuracy)

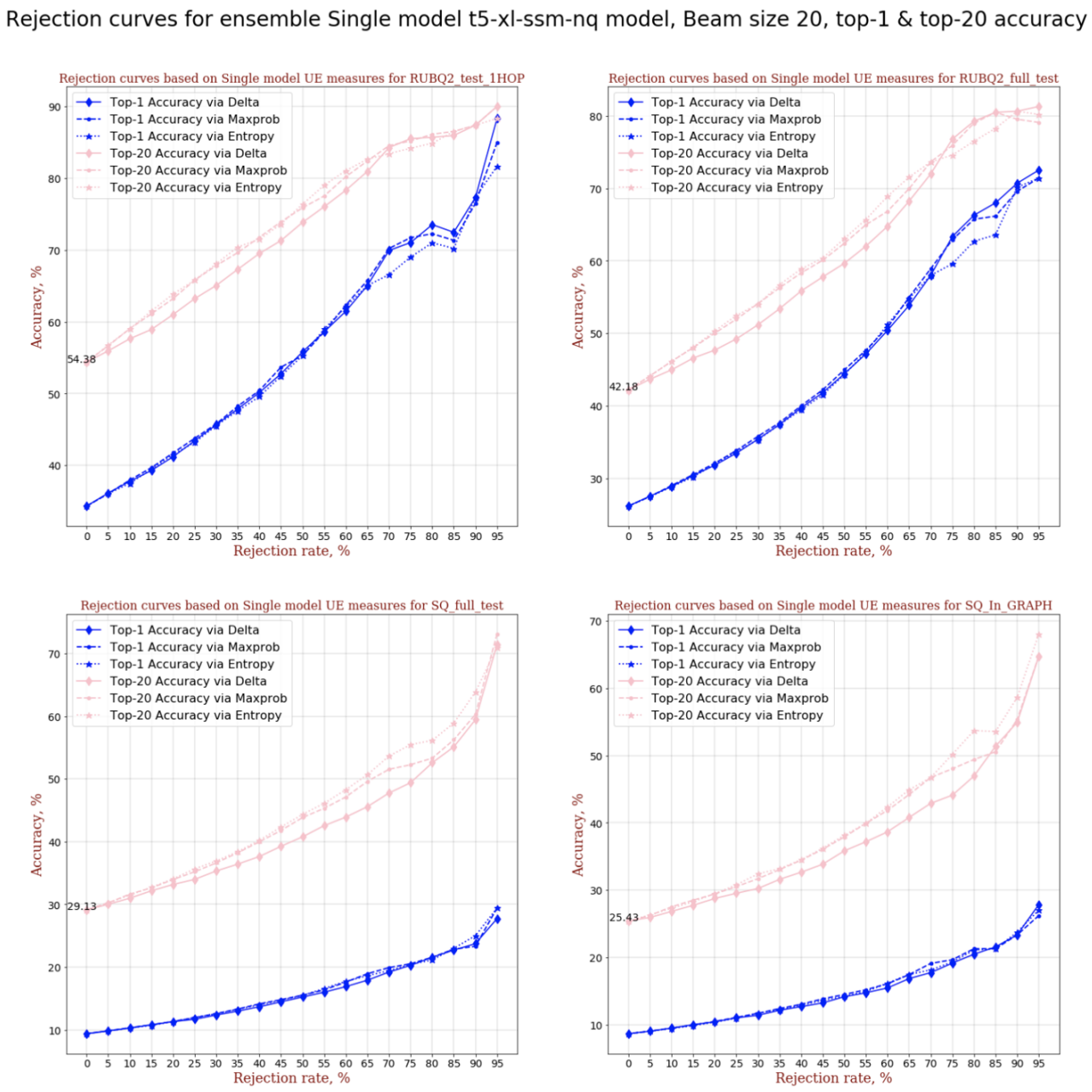

Figure 11. UE Results by single based model with beam size 20

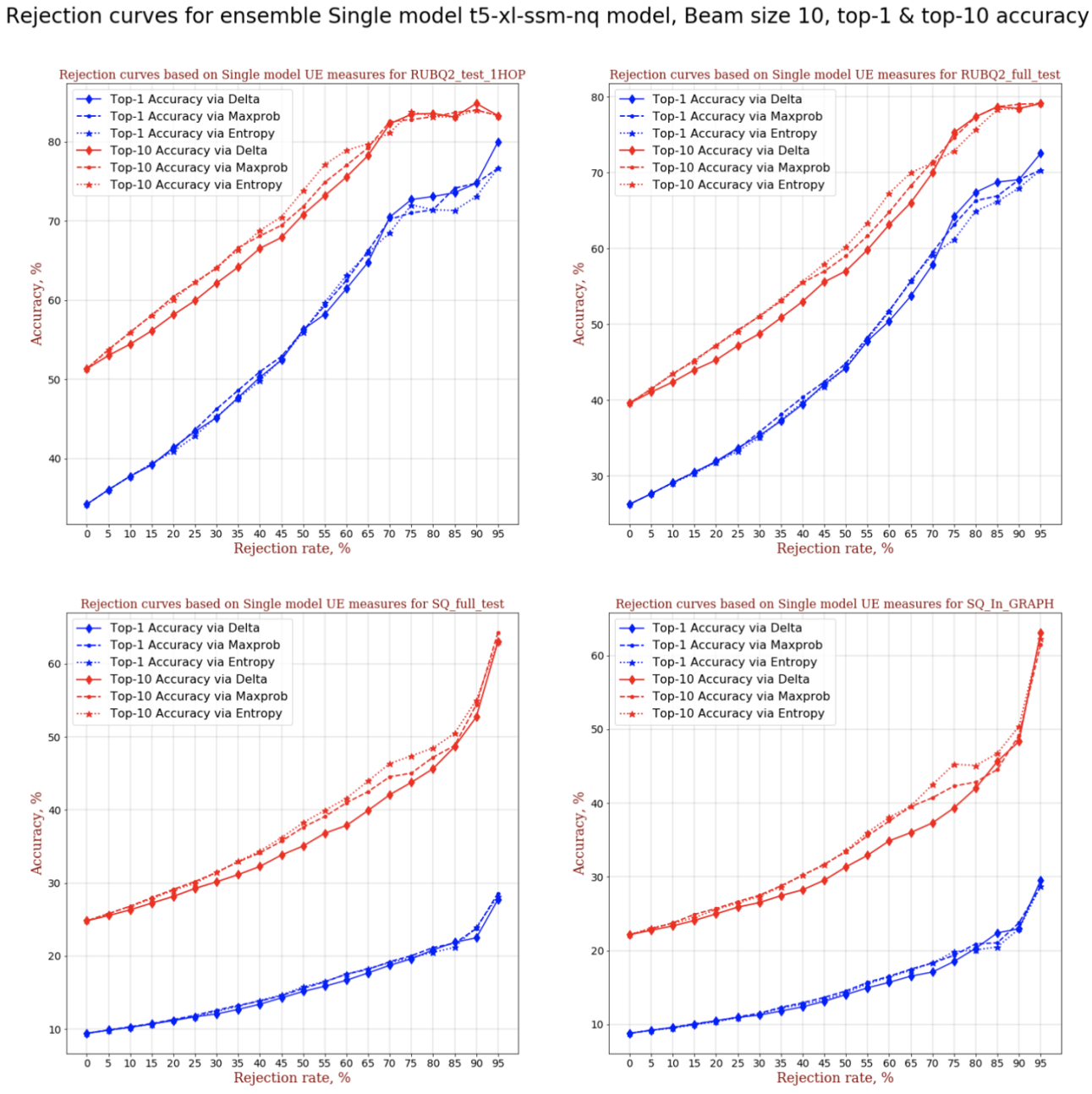

Figure 12. UE Results by single based model with beam size 10