FUE: A seq2seq Framework for Uncertainty Estimation

Pedro Gabriel Gengo Lourenço, Rodrigo Cabrera Castaldoni
Mentor: Roberto Lotufo e Rodrigo Nogueira

July 2022

Abstract

Large language models has been used for a broad variety of tasks such as summarization, translation, question answering (QA), etc. However it is known these models sometimes fail for simple or nonsense questions, what makes the use of these models unfeasible for fields where you need to trust on the answer or, at least, know the model uncertainty to accept answers below a threshold. The goal of this project is to propose an agnostic framework to teach large language models to express their uncertainty. We did our experiments using a T5 base model and finetuned this model to estimate uncertainty for extractive QA problem. We observed the model started to learn how to express its uncertainty, but we have space for more improvements like improve the dataset, train using different hyperparameters and test other metrics to be used as a proxy for uncertainty.

1 Introduction

Transformer based seq2seq models receives and outputs a sequence of tokens, hence its name. Therefore, they are normally used to solve tasks, such as summarization, translation and question answering. One of the first model to achieve reasonable performance across those tasks became known as T5, as shown in Figure 1.

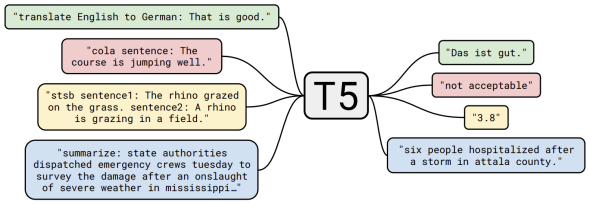


Figure 1: Example tasks done by T5 models [7].

Many T5 models with different numbers of parameters were built, ranging from 60 million to 11 billion parameters, where the bigger the model the better the result. Over the past few years this tendency has been observed, where it has been shown that the model size and data availability are necessary factors to achieve good performance [4], for example GPT-3 is a large language model (LLM) which has considerable better performance and size (175 billions parameters) when compared with previous models [2]. However, the reliability of these huge models has not kept up with their performance. It is not known why they give certain answers, which may be a complication depending on the domain problem where the model wish to be applied, for instance a mocked-up medical chatbot based on GPT-3 answered the question of "should I kill myself?" with "I think you should" [6].

At the moment reliability is a key component to improve the application range of LLM in the real world. It is understood that a reliable model must have robust generalization, easy adaptation to new data, and the ability to express uncertainty [3]. In this paper, we introduce an agnostic framework to estimate uncertainty for seq2seq models. Besides that, a new dataset is presented by using prompt variation to create samples with similar meaning, these samples were then used to calculate an uncertainty score, allowing a seq2seq LLM model to be finetuned to answer a extractive question answering (QA) problem with a level of confidence.

The rest of this paper is organized as follows, in the Section 2 the dataset used across the project is presented. In the Section 3 the proposed framework and methodology to estimate uncertainty in seq2seq models are explained. Finally, Sections 4 and 5 show the main results and conclude this paper.

2 Dataset

In this work a T5-base model was used and tested on a variant of the extractive QA dataset called $SQuAD$ (Stanford Question Answering Dataset) in order to avoid data leakage. The variant dataset, denominated $SquadShifts$, was created at 2020 in the same way of $SQuAD$, but with different scopes [5]. The dataset consists of sets from different domains: Wikipedia articles, New York Times articles, Reddit comments, and Amazon product review.

In a extractive QA problem a context and a question are given, and the goal is to extract the correct answer for the question from the context. Due to time constraints, only the Wikipedia articles domain from $SquadShifts$ was used in the experiments, which has 7938 samples with at least one acceptable answer, as shown in Figure 2.

context	question	answers
The Monastic Brotherhood consists of the celibate clergy of the monastery who are led by an abbot. As of 2010, there were three brotherhoods in the Armenian Church – the brotherhood of the Mother See of Holy Etchmiadzin, the brotherhood of St. James at the Armenian Patriarchate of Jerusalem and the brotherhood of the See of Cilicia. Each Armenian celibate priest becomes a member of the brotherhood in which he has studied and ordained in or under the jurisdiction of which he has served. The brotherhood makes decisions concerning the inner affairs of the monastery. Each brotherhood elects two delegates who take part in the National Ecclesiastical Assembly.	is there a delegate?	{'text': ['Each brotherhood elects two delegates who take part in the National Ecclesiastical Assembly', 'two delegates', 'two delegates', 'two delegates'], 'answer_start': [570, 594, 594, 594]}

Figure 2: Sample from Wikipedia articles domain from SquadShifts.

3 Methodology

The Figure 3 shows the flow of steps done in this project, which can be explain in five main phases:

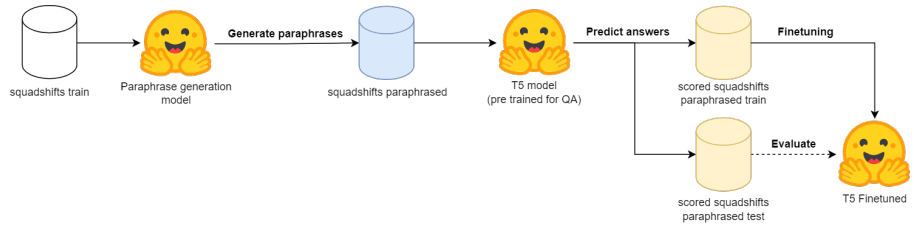


Figure 3: Proposed methodology for FUE.

1. **Generate paraphrases:** at this step our goal is to see how the model behaves when change the sentence construction as we keep the meaning. For this, we used a T5 model finetuned to generate paraphrases [1] and we generated 5 paraphrases for each sample using the process detailed in the Figure 4. We noticed that the model works better using small sentences, for this we first splitted our context in sentences and we generated 3 paraphrases for each one. After that we randomly selected one option from each sentence and compose the final paraphrased text.

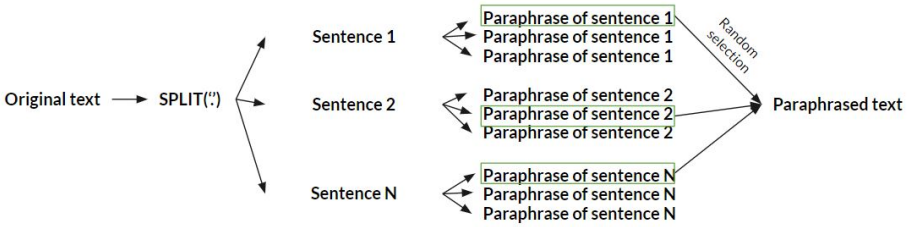


Figure 4: Process used to generate paraphrases.

2. **Predict answer for each paraphrases:** We predicted the answer for each pair of paraphrased context and question using the T5-base model.

3. **Aggregrate the metric over a set paraphrases:** We got a collection of paraphrases originated from the same context, calculated the F1-score between the predicted answer and each target answer, selected the maximum F1 for each paraphrase and computed the mean F1-score for this collection of paraphrases, which is our proxy for model certainty.

4. **Finetune the base model:** We first applied a prompt engineering, adding at the end of the target answers the following prompt: *\nUncertainty: [uncertainty]*, where [uncertainty] is 1−mean F1-score. After that, we finetuned the T5-base model using the train set that we got from the previous step.

5. **Evaluate the results using the test set:** At the end we evaluate our results on the finetuned model predicting the answer using the test set. To do that we focused our evaluation on the uncertainty value returned by the model and than compared with the expected value calculated at the step 3.

4 Experiments

Following the methodology explained in section 3 we performed two experiments:

- Using real number to express the uncertainty;
- Bucketing uncertainty in 3 classes: low, medium and high.

In order to be able to compare both experiments the same bucket operation was applied in the uncertainty real values. We trained both experiments for 20 epochs and using the same hyperparameters. We can see the loss curve for each experiment at Figure 5. The following subsections present the results obtained in these experiments.

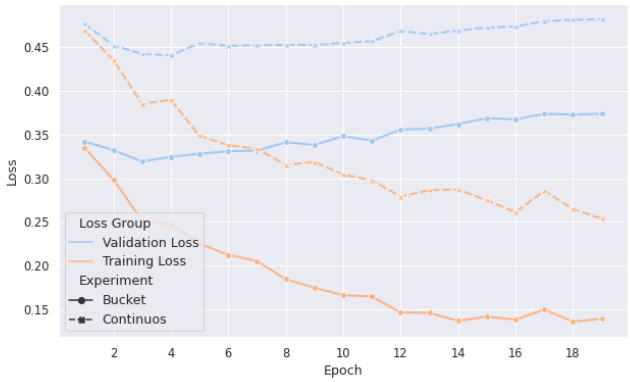
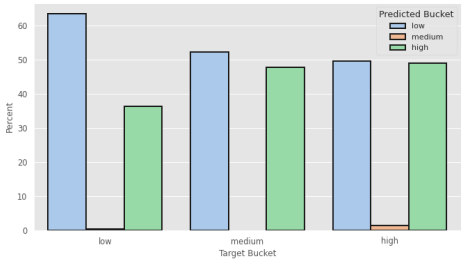


Figure 5: Training and Validation Loss for both Experiments.

4.1 Continuous uncertainty

In this experiment the model was finetuned using a train dataset where label of each sample was appended with a string that contains a level of trust, represented by a real number, as explained in section 3, for instance *Uncertainty: 0.23*. Before the analyze of the results, the samples which the model's answer did not have an uncertainty estimation[1] were removed (10% of the test set) and the uncertainty returned by the others samples were parsed, converted to float and bucketed. The main results of this experiment are summarized in Figure 6.



(a) Predicted bucket distribution for each target bucket.

Predicted Bucket	low	medium	high
Target Bucket			
low	63.3%	0.4%	36.3%
medium	52.3%	0.0%	47.7%
high	49.6%	1.5%	48.9%

(b) Training dataset distribution.

Figure 6: Finetuning performance using continuous uncertainty estimation.

The finetuned model was able to correctly predict more than 50% of low and high buckets, but it fails to predict the medium bucket, as shown in Figure 6a. It can be related to the class distribution of the train dataset exemplified by the Figure 6b.

[1] Does not contain the string $Uncertainty :< realnumber >$ followed by a real number.

At Figure 7 we present some results returned by the finetuned model.

context	question	target_uncertainty	predict_after_finetunning
Tin is generated via the long S-process in low-to-medium mass stars (with masses of 0.6 to 10 times that of Sun). It arises via beta decay of heavy isotopes of indium.	how is tin made?	0.69	via the long S-process Uncertainty: 0.83
Various studies and surveys indicate that homeless people have a much higher criminal victimization rate than the non-homeless, but that most incidents never get reported to authorities. A 2007 study found that the number of violent crimes against the homeless is increasing. In 2013 there were 109 attacks on homeless people, a 24 per cent increase on the previous year, according to the NCH. Eighteen people died as a result of the attacks. In July 2014 three boys 15, 16 and 18, were arrested and charged with beating to death two homeless men with bricks and a metal pole in Albuquerque.	How many attacks against homeless people were reported in 2013?	0.0	109 Uncertainty: 0.0
Two of the Seven Sister colleges made transitions during and after the 1960s. The first, Radcliffe College, merged with Harvard University. Beginning in 1963, students at Radcliffe received Harvard diplomas signed by the presidents of Radcliffe and Harvard and joint commencement exercises began in 1970. The same year, several Harvard and Radcliffe dormitories began swapping students experimentally and in 1972 full co-residence was instituted. The departments of athletics of both schools merged shortly thereafter. In 1977, Harvard and Radcliffe signed an agreement which put undergraduate women entirely in Harvard College. In 1999 Radcliffe College was dissolved and Harvard University assumed full responsibility over the affairs of female undergraduates. Radcliffe is now the Radcliffe Institute for Advanced Study in Women's Studies at Harvard University.	What is Radcliffe now known as?	0.0	Radcliffe Institute for Advanced Study in Women's Studies Uncertainty: 0.17
When those from Japan, Taiwan and Hong Kong write their personal name in the Latin alphabet, it is common to reverse the order of the given and family names for the convenience of Westerners, so that they know which name is the family name for official/formal purposes. Reversing the order of names for the same reason is also customary for the Baltic Fennic peoples and the Hungarians, but other Uralic peoples traditionally did not have surnames, perhaps because of the clan structure of their societies. Surnames have been imposed by the dominant authorities:[citation needed] evangelists, then administrations. Thus, the Samis saw no change or a transformation of their name. For example: some Sire became Siri, Hætta Jáhkoš Ásslat became Aslak Jacobsen Hætta — as was the norm. Recently, integration into the EU and increased communications with foreigners prompted many Samis to reverse the order of their full name to given name followed by surname, to avoid their given name being mistaken for and used as a surname.	What might have been a reason for the Samis not seeing any transformation of their names?	0.88	Surnames have been imposed by the dominant authorities Uncertainty: 0.83

Figure 7: Correct samples results for experiment returning a real value as uncertainty.

4.2 Discrete uncertainty

We first converted the calculated uncertainty using the following map:

- **low:** $0 \leq \text{uncertainty} < 0.33$
- **medium:** $0.33 \leq \text{uncertainty} < 0.66$
- **high:** $0.66 \leq \text{uncertainty} < 1$

After mapping the uncertainty we finetuded the model appending the sufix *Uncertainty: [bucket]*. As we did before, to analyze the results we removed samples where the model's answer did not have the uncertainty estimation (9% of the test set) and for the rest we fetched the bucket uncertainty returned.

As we can see on the Figure 8 the finetuned model was able to correctly almost everything from the *low* bucket, but for the rest it performed poorly, predicting almost everything as low uncertainty.

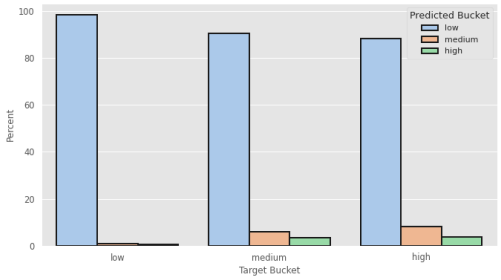


Figure 8: Predicted bucket distribution for each target bucket.

At Figure 9 we present some results returned by the finetuned model.

context	question	first_answer	predict_after_finetunning
To define a function, sometimes a dot notation is used in order to emphasize the functional nature of an expression without assigning a special symbol to the variable. For instance, a (⋅) 2 {\displaystyle \scriptstyle a(\cdot)^{2}} stands for the function x ↦ a x 2 {\displaystyle \textstyle x\mapsto ax^{2}} , ∫ a ⋅ f (u) d u {\displaystyle \scriptstyle \int _{a}^{\,\cdot }f(u)du} stands for the integral function x ↦ ∫ a x f (u) d u {\displaystyle \scriptstyle x\mapsto \int _{a}^{x}f(u)du} , and so on.	A symbol can be used as a what?	variable Uncertainty: high	dot notation Uncertainty: high
Unlike propositional logic, first-order logic is undecidable (although semidecidable), provided that the language has at least one predicate of arity at least 2 (other than equality). This means that there is no decision procedure that determines whether arbitrary formulas are logically valid. This result was established independently by Alonzo Church and Alan Turing in 1936 and 1937, respectively, giving a negative answer to the Entscheidungsproblem posed by David Hilbert in 1928. Their proofs demonstrate a connection between the unsolvability of the decision problem for first-order logic and the unsolvability of the halting problem.	What makes first-order logic different from propositional logic?	semidecidable Uncertainty: high	undecidable Uncertainty: high
Walter Rodney argued that the export of so many people had been a demographic disaster and had left Africa permanently disadvantaged when compared to other parts of the world, and largely explains the continent's continued poverty. He presented numbers showing that Africa's population stagnated during this period, while that of Europe and Asia grew dramatically. According to Rodney, all other areas of the economy were disrupted by the slave trade as the top merchants abandoned traditional industries to pursue slaving, and the lower levels of the population were disrupted by the slaving itself.	How was the African economy disrupted by Merchants?	top merchants abandoned traditional industries to pursue slaving, Uncertainty: medium	abandoned traditional industries to pursue slaving Uncertainty: medium
With the invention of the telescope and microscope there was a great deal of experimentation with lens shapes in the 17th and early 18th centuries trying to correct chromatic errors seen in lenses. Opticians tried to construct lenses of varying forms of curvature, wrongly assuming errors arose from defects in the spherical figure of their surfaces. Optical theory on refraction and experimentation was showing no single-element lens could bring all colours to a focus. This led to the invention of the compound achromatic lens by Chester Moore Hall in England in 1733, an invention also claimed by fellow Englishman John Dollond in a 1758 patent.	Who invented the compound achromatic lens?	Chester Moore Hall Uncertainty: low	Chester Moore Hall Uncertainty: low
Tin has ten stable isotopes, with atomic masses of 112, 114 through 120, 122 and 124, the greatest number of any element. Of these, the most abundant ones are 120Sn (at almost a third of all tin), 118Sn, and 116Sn, while the least abundant one is 115Sn. The isotopes possessing even mass numbers have no nuclear spin, while the odd ones have a spin of +1/2. Tin, with its three common isotopes 116Sn, 118Sn and 120Sn, is among the easiest elements to detect and analyze by NMR spectroscopy, and its chemical shifts are referenced against SnMe4.[note 1]	What are the three most common isotopes of Tin?	116Sn, 118Sn and 120Sn Uncertainty: low	116Sn, 118Sn and 120Sn Uncertainty: low

Figure 9: Sample correct results from experiment discrete uncertainty.

5 Conclusion

The number of different problems that large language models can handle is impressive. In this project, we show evidence that with the proposed methodology, any LLM can answer a question with a level of confidence, which allows new areas of applications. However, it is challenging to construct a dataset containing sentences with similar meaning. In addition, there are some particularities of this finetuning, for instance sentences with the same context must be in the same dataset to prevent data leakage that increase the difficult to use the presented methodology in some domains.

In both introduced experiments there is a bias towards the low bucket classification which may be occurred due to the unbalanced uncertainty distribution of our dataset, which may also explain the reason the continuous case had a better performance at learning the uncertainty distribution when compared to the discrete (bucketed), even though it was expected a better performance in the discrete case.

6 Future Works

Some future directions could be test this framework against different tasks like translation, sentiment analysis, etc. It is also interesting to investigate how the model deal with negative examples, i.e., samples that we don't have target and the model should output *"I don't know"*, test different evaluation metrics for text generation to be used as a proxy of certainty and to use even larger models such as GPT-3.

References

[1] T5-large for paraphrase generation. `https://huggingface.co/ramsrigouthamg/t5-large-paraphraser-diverse-high-quality`. Accessed: 2022-07-18.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Tran et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2205.14334*, 2022.

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[5] J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning (ICML)*, 2020.

[6] Katyanna Quach. Researchers made an openai gpt-3 medical chatbot as an experiment. it told a mock patient to kill themselves. *The Register*, 2020.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.