Multilingual Neural Question Generation

Thesis

Submitted in partial fulfillment of the requirements of
BITS F421T Thesis

By

Aditya Srikanth Veerubhotla

ID No. 2016A7TS0091H

Under the supervision of

Sandeep Aparajit,

Principal Software Engineer Manager, Microsoft

&

Dr. Aruna Malapati,

Department of CSIS, BITS Pilani Hyderabad Campus

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

HYDERABAD CAMPUS

(July, 2020)

Birla Institute of Technology & Science, Pilani
Hyderabad Campus

9th July, 2020

CERTIFICATE

This is to certify that the thesis entitled "**Multilingual Neural Question Generation** " submitted by **Veerubhotla Aditya Srikanth**, ID No **2016A7TS0091H** in partial fulfilment of the requirement of BITS **F421T** Thesis embodies the original work done by him/her under my supervision.

Signature of the Supervisor
Sandeep Aparajit,
Principal Software Engineer Manager,
Microsoft India.
Date: 9th July, 2020

Signature of the Co-supervisor
Dr. Aruna Malapati,
Associate Professor, CSIS,
BITS-Pilani Hyderabad Campus.
Date: 10 March 2020

Acknowledgments

I wish to express my sincere appreciation to my supervisors, Prof. Aruna Malapati and Sandeep Aparajit, for their valuable guidance and advice. I wish to express my deepest gratitude to Rajpal Kulhari and Rajarshee Mitra, for their mentorship. I would like to extend my wholehearted thanks to the People Also Ask team in Microsoft India, Hyderabad for their valuable insights and helpful discussions. I would like to express immense gratitude towards the management of BITS Pilani, Hyderabad Campus and Microsoft India, for offering me an opportunity to pursue an off-campus bachelor's thesis. Lastly, I thank my family and friends, who have stood by me and supported me in all my pursuits.

“It is not the answer that enlightens, but the question.”

Eugène Ionesco

Thesis Title: Multilingual Neural Question Generation
Supervisor: Sandeep Aparajit, Principal Engineer Manager, Microsoft
Session: Second Semester 2019-2020
Name of Student: Veerubhotla Aditya Srikanth
ID No: 2016A7TS0091H

Abstract

In this thesis, we present a detailed summary of the different approaches used for generating questions from a context passage that contains within it. We begin our investigation starting from Question Generation in English and develop techniques to improve the model's quality. We then investigate the importance of having a multilingual pretrained model for question generation, especially in cases where training data is scarce in a Zero-Shot paradigm. Finally, we show that current multilingual models suffer from Catastrophic Forgetting as they are fine-tuned on English, and experiment with different regularizers that can help in alleviating the problem.

Table of Contents

List of Tables

I. Introduction

Let us consider the following example: you are feeling unwell with an upset stomach and a fever. You go to a doctor seeking treatment and upon seeing you, the doctor asks a series of questions like have how long has it been since the pain started, have you eaten something outside, did you have any incidents of vomiting or diarrhea, etc. As you answer these questions, the doctor finally concludes his diagnosis: food poisoning, and prescribes medication to you which help you recover in a week.

The above example illustrates the importance of asking questions in exploration and understanding, with an underlying importance to asking the right questions to seek the information needed. Generating questions has a myriad of applications, from search wherein questions can be used to understand a user's information need, to education, where questions can be used as a tool to test understanding.

Another challenging and pertinent tasks in the field of Natural Language understanding and synthesis is the collection of massive amounts of task-specific data for training computationally expensive models, which is especially prohibitive for low resource languages. Current research tries to solve the problem by introducing multilingual models that are trained on a large multilingual corpus in an unsupervised manner with tasks that help in understanding the relationship between different languages. These models are then fine-tuned on a downstream task in a language for which task-specific data exists and infer on all languages. This "Zero-Shot Inference" on the downstream tasks have shown promise in many discriminative tasks and have much lower requirements for labelled training data.

In this dissertation, we aim to create a multilingual Text-Based Question Generation (mQG) system that takes a passage of text in a given language (e.g. an article from Wikipedia) and generate a relevant, well-formed question that is answered by the passage in the same language.

A. Task Definition

Formally, given a passage P in a language L, Multilingual Question Generation (mQG) aims to generate a grammatically correct question Q in language L that is answered by the passage P. For example, Given the English passage,

> *The onset of flu often seems sudden: people describe feeling like they've "been hit by a truck.". Common flu symptoms include sudden onset, fever and chills, cough, muscle and joint pain, headache, fatigue and weakness. Some people also get a stuffy nose and sore throat.*

the task is to generate a question,

> *What are the symptoms before the onset of a fever?*

B. Types of Questions

1. Based on structure

a) Yes/No Questions

The questions that require the response to be either an affirmation or a refutation come under this category. Examples include “Is the sky blue?”, “Is it going to rain today?”, “Will we have a crewed landing on Mars this year?”, etc. An extension of this category includes Multiple Choice Questions, wherein the answer is a selection of an answer from given alternatives. For example, int the question “Are you going to support India or England in the upcoming cricket match?”, the response to these are either the affirmative answers: “India,” “England,” “Both,” or the dismissive: “Neither.”

b) Wh-Questions

Wh-questions use special interrogative words to request for specific information including location (Where is the Eiffel Tower?), definition (What does photosynthesis?), identity of a person (Who was Mahatma Gandhi?), causation of an event (Why did WW2 happen?), time of occurrence of an event (When did the Titanic sink?), etc.

c) Fill-in-the-blank Questions

These questions are a case of *Cloze Tasks* (Taylor, 1953), where certain words or phrases are removed, and the answerer is required to complete the sentence. For example:

The state of water at 30^{o}C is a _____.

d) Tag Questions

These questions are formed by adding an interrogative phrase (called a question tag or question tail) into a declarative or imperative statement. Example:

It's hot out there, isn't it?

e) Indirect Questions

Indirect questions (also called *interrogative content clauses*) are used as subordinate clauses in sentences and emphasize knowledge or lack of knowledge of an element of a fact. Example:

I wonder where Ram and Harish went.

2. Based on subjectivity

a) Factual

Factual Questions are questions based on information from a widely accepted, verifiable resource. Example:

Does the Earth revolve around the Sun?

b) Subjective

Subjective Questions that encourage the answerer to express his/her opinion and provide an explanation.

What do you think I should do to ace this thesis?

3. Based on descriptiveness

a) Brief

The expected answer to these questions is succinct and do not elaborate on many ideas. Example:

What is the state capital of Minnesota?

b) Elaborate

The answers are elaborate and cover many ideas/themes. Example:

Describe the significant events of the Second World War.

C. Multilingualism

While there have been many advances in the field of Natural Language Processing (NLP) and Natural Language Generation (NLG), they are predominantly restricted to English, catering to about 13 percent of the population[1]. Hence, to democratize the availability of cutting-edge technology to multiple languages, efforts on building multilingual systems should be encouraged.

An important distinction needs to be drawn between techniques that are *multilingual* and *cross lingual*, which are often conflated together but represent different ideas. *Multilingual* systems generate outputs that are in the *same* language as the input, whereas in the case of *Cross Lingual* systems, the output language *need not be the same* as the input language.

D. Applications of Multilingual Question Generation

Answering questions plays a pivotal role in understanding and exploration. In learning, question answering is a crucial didactic tool that allows evaluation through reflection. In scientific or literary exploration, questions help in grounding ideas and in creating opportunities for inquiry.

[1]www.ethnologue.com

In education, QG plays the role of an evaluator of understanding. It is common to find questions at the end of a chapter in a textbook, designed to test the understanding of a person on the concepts taught in a chapter. An application of QG could be in the creation of conversational quizzing agents that generate challenging questions that enforce clarity in understanding for answering a question. Questions also provide a summary of the topics being taught in a chapter, allowing the student to isolate the information and learn what is required and revise what is necessary.

An exciting application of QG is in search. Generation of right questions assists the search engine to arrive at the right answers needed by users to satisfy their information need, especially when the required information is abstract. An example of this is a doctor (search engine) and a patient (user). Patients present their symptoms to the doctor, who asks a series of questions that help in the diagnosis of the illness, subsequently providing medication (the required information).

E. Problem Statement

The aim of the thesis is to build automatic multilingual Question Generation Systems that take a natural language passage as an input and generate brief, factual questions of various structural types and in the same language as the passage.

F. Contributions of the thesis

1) We develop an end-to-end question generation system capable of generating brief factual questions from an input passage.
2) We evaluate currently available multilingual generative neural models on QGen and evaluate the performance in supervised and zero-shot settings.

G. Thesis organization

This thesis is organized into multiple Chapters. In Chapter 2, we provide a literature review of the previous approaches on question generation and developing multilingual models. In Chapter 3, we

describe our approach, providing details on the pipeline created, with the data used and models experimented with. In Chapter 3, we present our results and providing an analysis of the results in Chapter 4. The subsequent Chapters rests with the conclusion of the thesis and future work.

II. Literature Review

A. Question Generation

The generation of factual interrogative sentences has seen much interest over the years. The initial approaches relied on designing hand-crafted transformations based on patterns from the data being used (Ross, 1967; Wolfe, 1976; M. A. Walker, 2001; A. C. Graesser, 2005; Heilman, 2009; Rus, 2010). While these approaches showed high precision, the curation of transformation rules is expensive and time-consuming. Moreover, Vanderwende (2008) showed that in order to ask good questions, more sophisticated methods than a syntactic transformation of a declarative sentence are needed. Hence, further research focused on Machine Learning based approaches, which allowed for context-sensitive transformations to be learned from the data. Machine Learning based approaches can be broadly classified into two categories: those that viewed QG as a Sequence to Sequence (Seq2Seq) task, or those that viewed it as a language modeling (LM) task.

Seq2Seq (Sutskever, 2014) architectures decompose generation into two steps: obtaining the representations of the inputs and then utilizing this representation to generate outputs. Many Recurrent Neural Network (RNN) + attention (Cho, 2014; Bahdanau, 2015) based Seq2Seq architectures have been proposed for QG (Du, 2017; Song, 2018; Hosking, 2019; Kim, 2019; Li, 2019), which employ attention mechanism (Cho, 2014; Bahdanau, 2015) that allows the model to focus on relevant information while generating information. Another set of approaches use pretrained Transformer (Vaswani, 2017) based architectures and has significantly outperformed previous models (Bao, 2020; Dong, 2019).

Another way of formulating QG is to treat it as a variant of a Language Modeling task, where the question is appended to the answer with a special symbol separating the passage and the question. During training, the model predicts (decodes) the next token conditioned on the answer and the previously generated question tokens. This can be an easy way to fine-tune pretrained decoder models such as OpenAI's GPT-2 (Radford, 2019).

Additionally, generation techniques like Copy Mechanism (Gu, 2016; Gülçehre, 2016) and Beam Search have been used to improve the quality of the questions generated. Different learning paradigms, such as Adversarial Learning (Hosking, 2019) and Reinforcement Learning (Hosking, 2019), have been used to explicitly induce task-specific training.

B. Multilingual Language Models

Training models on extensive amounts of data (e.g. CommonCrawl[2] or Wikipedia[3]) in a self-supervised manner has allowed the models develop representations for syntactic and semantic structure. Many different training techniques have been proposed for pre-training, which can be broadly grouped into: Denoising Autoencoding (DAE) and Causal Language Modeling (CLM). In denoising autoencoding, the inputs are corrupted using a noising function (omitting or reordering phrases, shuffling sentences, etc.), and the model learns to predict the original sequence (Peters, 2018; Devlin, 2019; Lample, 2019; Huang, 2019; Liu, 2020; Liang, 2020). In Causal Language Modeling, the model learns to predict the next word conditioned on previous predictions (Lample, 2019; Radford, 2019). This pretraining has shown large improvements over a large range of understanding and generation tasks, and not requiring a large dataset in doing so.

An interesting ability of such multilingual and cross-lingual pretrained models is that they can predict on languages it has not been trained on, termed as Zero-Shot Learning (ZSL). ZSL is an economically important paradigm for both academia and industry, since it requires training data to be curated in a few high resource languages like English, French and German, and then inferred on multiple languages including low-resources like Hindi, Urdu, and Telugu. While Zero-Shot Learning has been applied for Natural Language Understanding (Huang, 2019; Lample, 2019; Liang, 2020; Conneau, 2020), it is only recently that researchers are looking into the harder problem of generation (Liang, 2020). Finally, as more research is conducted into this field, we could hope for a unified multilingual model that can transfer task knowledge across multiple languages and requiring fewer examples to do so.

[2] https://commoncrawl.org

[3] https://www.wikipedia.org/

III. Methods

In our attempt to develop a multilingual question generation model, started with English, since it allowed us to understand the nuances of the problem. Subsequently, we expanded to two other languages namely French and German. Finally, we attempt to create a model capable of Zero-Shot generation. The hyperparameters used for fine tuning the models have been mentioned in *Table 1* Hyperparameters used All models were trained on 4 Nvidia V100 32GB GPUs with FP16 precision.

A. QG in English

We began our experimentation in English using GPT-2 medium (Radford, 2019), a 24-Layer, 345M parameter Transformer decoder trained on CLM on Wikipedia text. The model was trained on a filtered and lower-cased dataset containing one million Question and Passage (QP) pairs. The input to the model was of the form

Passage [SEP] *Question* [SEP]

Where "[SEP]" was a special token added to the vocabulary that was used as a separator between the Passage and the Question and also as an End-Of-Sequence (EOS) tag. The model was trained using CLM objective with Cross Entropy (Xent) as the loss function. During training, the loss computed over prediction of the passage was masked out, so that the model learns to predict the generation of a question.

While the model was able to give good results, the size of the model hindered it being shipped into production. Subsequently, we applied Knowledge Distillation (Appendix A) to reduce the number of layers from 24 layers to 6, while retaining as much of the quality as possible, allowing it to be sent for production. Subsequent evaluations suggested that the model suffered from generating questions that were weakly relevant or irrelevant to the passages. This was frequent especially when the passage had multiple named entities or was expressing multiple ideas in one sentence.

To induce information of relevance in the model, inspired by the work done in (Zhang, 2019) and (Rennie, 2017), we follow the Self Critical Sequence Training (SCST) which is a form of the popular REINFORCE algorithm that, rather than estimating a "baseline" to normalize the rewards and reduce variance, utilizes the output of its own test-time inference algorithm to normalize the rewards it experiences (Rennie, 2017). For the Self Critical loss, we consider two rewards: a Question Answering (QA) model that predicted whether a given passage could answer a given question and ROUGE-L gain as, as described in (Rennie, 2017). The distilled model was first trained using Xent as discussed above for a few epochs, then the model was further trained on a loss that interpolated between the rewards and the Xent loss. The objective function could be summarized as:

$$L = \begin{cases} L_{\text{Xent}}, & \text{if } n < N. \\ \gamma * L_{\text{Xent}} + (1-\gamma) * L_{\text{rl}}, & \text{otherwise.} \end{cases}$$

$$\nabla L_{\text{rl}} = -E_{w^s \sim p_\theta}[(r_{total}(w^s) - b)\nabla_\theta \log p_\theta(w^s)]$$

$$\text{L}_{\text{Xent}} = \sum_n \sum_c \hat{y}_n \log p(C_c | x_n, T)$$

$$\text{r}_{total} = r_{QA} + r_{ROUGE-L}$$

Where γ and N are Hyperparameters. In our experiments, we used γ = 0.25, N = 2.

B. Extending to other languages

Having understood the main challenges in QG, we began our investigation in mQG. Here we conducted experiments in two broad settings: 1. Supervised, where we assume enough data is available to train the model which is used primarily for evaluating models, and 2. Few-Shot, in which the availability of labelled data is scarce and expensive to curate, modeling real life scenarios.

1. A Supervised Setting

We begin our experiments with an (unrealistic) assumption that there is enough labelled available to be able to be able to generalize well. This is done primarily to investigate the asymptotic performance of the models as a function of the data. We conduct experiments focused on three languages: English, French and German. As a baseline, we work with a model that has been pretrained on English, subsequently multilingual models. We curate a dataset from a cache containing QP pairs, apply deduplication, pass it through a rule-based classifier for filtering out non-questions, and lowercasing the data, obtaining a dataset of 1M QP pairs each.

To establish a baseline, we conduct experiments with GPT-2 since the authors have shown that the model has the capability of translation owing to its extensive pretraining. We begin our experiments with the English RL model and train it directly on the French and German dataset, to evaluate its ability to transfer the knowledge to other languages. We then experiment with GPT-2 Medium and continue training as described in *III (A).* We then posit that we could improve the generation and avoid language intermixing between different languages by creating language specific vocabularies and utilize these specific vocabularies for generation. We curate our vocabularies by sampling 1M Wikipedia articles and running Byte Pair Encoding (Sennrich, 2016), and restrict the vocabulary to fifty thousand in each language, with common tokens between any two languages having the same embeddings. We then fine-tune the model by oversampling French and German, to allow the vocabulary tokens to be learned.

Having set a baseline, we then investigated pretrained multilingual models. We choose mBART (Liu, 2020), and Cross Lingual Future N-Gram Prediction model (henceforth called xProphetNet or xPN) described in XGLUE (Liang, 2020) as they demonstrated good performance across different tasks and by being Seq2Seq models, they generate better contextual encodings for an input sentence than Autoregressive methods (Raffel, 2019). The models are trained on the same dataset, and we minimize the Xent loss over the questions generated. For mBART, the model requires the presence of a language identifier (LID) for which Fasttext's language identification

tool[4] was used to identify the source language. We then experimented with different sizes of the dataset for fine-tuning and found that a dataset size greater than 200,000 examples in each language gave good results and a decrease in number of pairs showed a drop in quality when judged.

In conclusion, we find that current models can make good predictions given enough training data is available. However, in most real-life scenarios, it is difficult to obtain labelled data in multiple languages, making it challenging for developing high-quality solutions incorporating these models. In the next section, we investigate whether pretrained models can generate questions in languages for which data is scarce, or in an extreme case, unavailable.

2. Zero-Shot Cross-Lingual transfer: Are we there yet?

In this section we conduct experiments to answer two questions: which of the above pretrained models display the ability of zero-shot cross lingual transfer and how do we enhance this ability. We curate a set of 500,000 QP pairs as the training data, with test data taken from xGLUE and TyDi QA (Clark, 2020). We fine-tune mBART and xPN on this data and evaluate on all languages, running training for two epochs and storing checkpoints for every 5000 updates. From the results shown in *Error! Reference source not found.*, we observe that mBART does not show the ability of zero-shot transfer, while xPN does, hence we conduct all further experiments with xPN.

We formulate the cross-lingual transfer as a knowledge-retention problem, where the pre-trained model possesses great understanding of the syntax and semantics of multiple languages, but loses its multilingual ability while being trained on a monolingual dataset. To retain the knowledge while simultaneously adapting to a new task, some sort of constraint is required. We experiment with different regularization techniques commonly applied in literature including layer freezing where we stop the gradients to the first six layers of the encoder and the last six layers of the decoder, usage of weight decay, and increasing Dropout (Srivastava, 2014) to constrain updates.

[4] https://fasttext.cc/docs/en/language-identification.html

We finally add a few examples extracted from websites containing Frequently Asked Questions (FAQs) and fine-tune the model on them.

Model	Batch Size	Learning Rate	Adam Beta1	Adam Beta2	Dropout	Weight Decay
GPT-2	32	5e-05	0.9	0.999	0.3	0.0
mBART	32	3e-05	0.9	0.98	0.3	0.0
xPN	16	1e-5	0.9	0.999	0.1	0.01

Table 1 Hyperparameters used

IV. Results and Discussion

In this chapter, we present the results of the experiments conducted, and some insights that we obtained from them. Metrics used for evaluation are BLEU (Papineni, 2002) which measures the n-gram precision between the source and target, and ROUGE-L, which measures recall (Lin, 2004), which is a recall oriented metric.

A. QG in English

Model	BLEU-4	ROUGE-L
GPT-2 Medium	0.24	0.56
Distilled GPT-2	0.21	0.53
Distilled GPT-2 + RL	**0.25**	**0.59**

Table 2 Results on English

As shown in Table 2 Results on English, the Reinforcement Learning fine-tuned model significantly outperformed all previous English models to obtain good scores in English. In Table 3 Comparison between Distilled and RL + Distilled model., we compare the improvements that Reinforcement Learning brought in English. The QA reward allowed the model to learn how to generate questions that are relevant to the passage while ROUGE gain acts as a proxy for the question's grammar: the higher the ROUGE, the better the question.

Passage	Before	After
Ritwik married Surama Devi, niece of active left- wing member, Sadhana Roychowdhury. They separated and she went to her ancestral place, Shillong after Ritwik was temporarily sent to a mental hospital.	*who is "ritwik s" niece*	*who is ritwik married to*
Ghatak entered the film industry with Nimai Ghosh's Chinnamul (1950) as actor and assist ant director. Chinnamul was followed in two y ears by Ghatak's first completed film Nagarik (1952), both major breakthroughs for the India n cinema.	*what was the first film of chinnamul*	*who was ghatak in the film industry*

sebaceous glands (oil glands) small glands loc ated within the epidermis, and associated with the hair follicle, that produce and secrete an oily substance that lubricates the skin and hair.	*what are the glands in the skin*	*where are sebaceous glands located*
since 2001, the president has earned a $400,00 0 annual salary, along with a $50,000 annual expense account, a $100,000 nontaxable travel account, and $19,000 for entertainment. en.wikipedia.org	*how much does an en wikipedia employee make*	*how much money does president get paid*

Table 3 Comparison between Distilled and RL + Distilled model.

B. mQG

1. Supervised

Model	Metric	English	French	German
Distilled GPT-2 + RL	BLEU-4	*0.03*	*0.08*	*0.06*
	ROUGE-L	*0.24*	*0.37*	*0.35*
GPT-2 Medium	BLEU-4	0.20	0.12	0.11
	ROUGE-L	0.51	0.43	0.41
Distilled GPT-2 + Language-Specific Vocabulary	BLEU-4	0.20	0.13	0.07
	ROUGE-L	0.506	0.44	0.36
xPN	BLEU-4	0.21	0.16	0.136
	ROUGE-L	0.54	0.46	0.43
mBART	BLEU-4	**0.22**	**0.17**	**0.137**
	ROUGE-L	**0.54**	**0.47**	**0.44**

Table 4 Results in English, French and German

From

xPN	BLEU-4	0.21	0.16	0.136
	ROUGE-L	0.54	0.46	0.43
mBART	BLEU-4	**0.22**	**0.17**	**0.137**
	ROUGE-L	**0.54**	**0.47**	**0.44**

Table 4 Results in English, French and German, we see that the fine-tuned model has overly specialized for the task of English QG that it has lost its ability of multilingual generation upon being fine-tuned on French and German data. This is an example of "Catastrophic Forgetting", where the model forgets its previous knowledge when it sees new data. Second, curation of a language specific vocabulary allowed the model to get significantly better results, nearly matching the performance of a model four times its size. This could be attributed to the intermediate layers learning language agnostic semantic information, while the initial and final layers learn the syntax.

Additionally, we observe that the multilingual pretrained models obtain achieve better scores overall, gains obtained from the models are not as large as one would expect. This indicates that given enough supervised data, multilingual and monolingual models achieve nearly equivalent results. We also observe that while the rresults for xPN and mBART are close to each other, the scores obtained by mBART outperform those obtained by xPN. This could be possible ascribed to a phenomenon described in (Conneau, 2020) called Curse of Multilinguality, where the performance gains obtained for a given model decreases as the number of languages on which it is pre-training on increases. This is evident in the case of xPN which is trained on 100 languages versus mBART which is trained on 25 languages.

2. Few-Shot

Model	Metric	Italian	Russian
xPN	ROUGE-L	0.12	0.08
	BLEU-4	0.008	0.003
xPN + Dropout 0.3	ROUGE-L	0.11	0.08
	BLEU-4	0.008	0.002
xPN + Dropout 0.3 + No weight decay	ROUGE-L	0.13	0.08
	BLEU-4	0.011	0.002
xPN + layer freezing	ROUGE-L	*0.12*	*0.10*
	BLEU-4	*0.007*	*0.002*
xPN +	ROUGE-L	**0.35**	**0.15**

FAQ data (10k)	BLEU-4	**0.13**	**0.02**

Table 5 Results on zero-shot and few-shot settings

From Table 5 Results on zero-shot and few-shot settings it is evident that the presence of a few examples improves the performance of the models. However, as the table shows, the gains achieved in Italian are greater than those in Russian. This can be attributed to the presence of greater number of tokens between English and Italian, which share the same script, over English and Russian, which virtually don't have any common tokens. Furthermore, across all regularization techniques, layer freezing shows the highest scores, this can be explained by the fact that updates are prevented, which retains the model's pretraining knowledge. The results also suggest that the usage of weight decay is detrimental to the generalization performance, this could be explained by the fact that the regularization enforces the weights to drop to zeros, drifting away from their original values.

xPN + Layer Freezing	Metric	Italian	Russian
Learning rate = 5e-05	ROUGE-L	0.12	0.10
	BLEU-4	0.007	0.002
Learning rate = 1e-07	ROUGE-L	**0.22**	**0.14**
	BLEU-4	**0.044**	**0.03**

Table 6 Effect of Learning Rate

Finally, there seems to be a large effect of learning rate in the generalization ability of the model, with the presence of a higher learning rate, the model aggressively pursues learning English QG, causing Catastrophic Forgetting. This underscores the importance of choosing the right set of hyperparameters for not allowing the model to "overfit" on the language and retain the pretraining knowledge.

V. Conclusion

We provided a detailed analysis on the current state of multilingual question generation, and our experiments unearthed a few crucial questions: how do we encourage pretraining such that the models learn language representations without a large dependence on token level representations, and, the need for building regularization techniques so that Catastrophic Forgetting could be avoided. If these two questions could be addressed and improved upon in the future, we can truly expect multilingual systems capable of Zero-Shot cross lingual task adaptation.

VI. Appendix

A. Knowledge Distillation

1. The concept of distillation

Knowledge Distillation (Hinton, 2015) aims to transfer task specific knowledge of large computationally expensive models (often ensembles) to smaller models without much loss in performance. The framework consists of two agents: a computationally expensive, well-generalized teacher model, and a smaller student model. The teacher model generates "soft targets", which are probability distributions that the student optimizes over along with "hard targets" which is obtained from the dataset. The idea is to teach the student what the teacher has learned through the soft targets, which represent the knowledge accrued by the teacher during its training.

2. Formulation

Neural Networks applied to classification tasks produce the probabilities of each class by applying a Softmax operation over the generated logits. These can be represented as

$$\mathrm{p}(\mathrm{C}_i|x,T) = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)}$$

Where $p(C_i|x)$ is the probability of the class C_i, z_i is the logit generated for the class, T is called temperature which is used to soften the probability distribution. The cross-entropy (Xent) sloss can be expressed as:

$$\mathrm{L}_{CE} = \sum_n \sum_c \hat{y}_n \log p(C_c|x_n,T)$$

Where the loss is computed over the dataset N having C classes with the class label being $\hat{y}_n$ is the true label.

To allow the transfer of knowledge between the student and the teacher, the student model tries to mimic the teacher distribution, and this can be expressed as the Kullback–Leibler (KL) divergence between the student and the teacher distributions.

$$\mathrm{L}_{KL} = \sum_n \sum_c q(C_c|x_n, T) \log p(C_c|x_n, T)$$

Finally, the distillation loss is the interpolation between the Xent loss and the KL divergence loss.

$$\mathrm{L}_{distill} = \gamma L_{CE} + T^2(1-\gamma)L_{KL}$$

Where T^2 is multiplied to the KL divergence loss so as to normalize the inverse square of the temperature generated in the gradient.

3. Distillation in Question Generation

We applied the concepts described above to distill the 24-layer GPT-2 model into a 6-layer model. We sampled 500k QP pairs in English, set gamma and Temperature as 1, and computed the above loss for each time-step of the model's prediction. The results obtained are mentioned below:

Model	BLEU-4	ROUGE-L
GPT-2 Medium	0.24	0.56
GPT-2 Distill	0.21	0.53

Table 7 Results of Distillation

As we can see, the distilled model shows comparable performance to that of the teacher model while being four times faster.

B. Examples

In this section, we present a few examples of the questions generated by models we experimented with.

Passage	**Question**	**Translated passage**	**Translated question**
Mit der Löschung wegen Vermögenslosigkeit entfallen die Notarkosten für die endgültige Löschung der GmbH, Publizitätspflichten entfallen und nachfolgend ist auch kein Jahresabschluss mehr durchzuführen, was in der Regel auch das Hinzuziehen eines Steuerberaters bedeut et, der für seine Arbeit entlohnt werden möchte	*ist eine loeschung der gmbh ein jahresabsch*	**With the cancellation due to lack of assets, the notary costs for the final deletion of the GmbH are eliminated, publicity obligations are eliminated and subsequently no annual financial statements must be carried out, which usually also means the use of a tax advisor who wants to be paid for his work.**	**is a loeschung of gmbh a year-round**
Das Kugelstoßen (Englisch: [&shot put)&] gehört zu den Wurfdisziplinen der Leichtathletik und als Teildisziplin zu m Siebenkampf sowie zum Zehnkampf. Eine Metallkugel muss durch impulsartiges Wegschleudern mit dem Arm s o weit als möglich gestoßen werden.	*was ist ein kugelstoss*	**The shot put is one of the throwing disciplines of athletics and as a sub-discipline for heptathlon as well as decathlon. A metal ball must be pushed as far as possible by impulse-like hurling away with the arm.**	**what is a shot put**
Ramasuri Country. Ramasuri Country-Time ist Samstags von 18 bis 20 Uhr. Radio Ramasuri zäh lt zu den wenigen europäischen Radiostationen, die Mitg lied sowohl der amerikanischen als auch der kanadischen Country Music Association sind.	*was ist ramasuri country*	**Ramasuri Country. Ramasuri Country-Time is Saturdays from 6 p.m. to 8 p.m. Radio Ramasuri is one of the few European radio stations that are members of both the American and Canadian Country Music Associations.**	**what is ramasuri country**

Table 8 Multilingual GPT-2 Predictions in German

Passage	Question	Translated passage	Translated question
Récemment, l'accent a été mis sur les intolérances aux sucres comme le fructose et le sorbitol. La diarrhée et l'inconfort abdominal chronique sont des signes classiques de malabsorption du fructose et/ou du sorbitol.	*pourquoi je me rejete trop souvent*	**Recently, the focus has been on intolerance to sugars like fructose and sorbitol. Diarrhea and chronic abdominal discomfort are classic signs of fructose and / or sorbitol malabsorption.**	**why i reject myself too often**
Or les populations des pays en développement [&n'ont pas d'accès à l'eau potable&] et à l'assainissement : elles ne peuvent et ne savent pas toujours comment faire pour éviter de polluer l'eau et de tomber malades. Ainsi, des populations tombent malades et meurent à cause du manque d'information et d'installations sanitaires.	*pourquoi les pays ont pas d'eau potable.*	**People in developing countries [& have no access to clean water &] and sanitation: they cannot and do not always know how to avoid polluting the water and falling ill. As a result, people fall ill and die due to the lack of information and sanitation.**	**why countries have no drinking water.**
La communication non verbale. La communication non verbale a une grande importance dans la vie sociale de l'individu. Et suivant des recherches, elle représente plus de 75 % de la communication. Fondamentalement, il est l'un des aspects clés de la communication (et d'autant plus important en culture de haute contexte).	*qu'est ce que la communication non verbale*	**Non-verbal communication. Nonverbal communication is of great importance in the social life of the individual. And according to research, it represents more than 75% of communication. Basically, it is one of the key aspects of communication (and even more important in high context culture).**	**what is nonverbal communication**

Table 9 Multilingual GPT-2 Predictions in French

Passage	Question	Translated passage	Translated question
Mehmet ist eine Abwandlung von Muhammed (nach dem Propheten Mohammed) im Türkischen. Er bedeutet ursprünglich „der Gepriesene" oder „der Gelobte". Diese Namensform war bereits durch das Osmanische Reich in Europa bekannt.	*was heisst emmanuel*	**Mehmet is a variation of Muhammed (after the Prophet Mohammed) in Turkish. It originally means "the praised" or "the promised". This form of name was already known in Europe through the Ottoman Empire.**	**what does emmanuel mean**
Handwerkerleistungen werden in der Steuererklärung au f Seite 3 des Hauptvordrucks unter dem Punkt "Haushalts nahe Beschäftigungen, Dienstleistungen, Handwerkerleis tungen" in Zeile 75 eingetragen.	*wo werden handwerkerleistungen in der steue rerklaerung eingetragen*	**Craftsman services are entered in the tax return on page 3 of the main form under "Household employment, services, craftsman services" in line 75.**	**Where craftsman services are entered in the tax return**

Messi lebt in einer kleinen Gegend namens Bellamar, ein em prestigeträchtigen und teuren Vorort von Castelldefel s, Barcelona.	*wo lebt messi in barcelona*	**Messi lives in a small area called Bellamar, a prestigious and expensive suburb of Castelldefels, Barcelona.**	**Where does messi live in Barcelona**

Table 10 mBART Predictions in German

Passage	Question	Translated passage	Translated question
"Choix d'une TV : Les différents critères. OLED, UHD, Full HD, LED, LCD, 4K, HDR, QLED ... Nous allons re mettre de l'ordre dans vos idées pour que vous compreni ez bien les différentes caractéristiques techniques des télé viseurs.	*quels sont les differents types de tv*	**"Choice of a TV: The different criteria. OLED, UHD, Full HD, LED, LCD, 4K, HDR, QLED ... We will put your ideas in order so that you understand the different technical characteristics of the iseurs "**	**what are the different types of tv**
Connectez votre iPhone en utilisant le câble USB. Dès qu e votre appareil aura été reconnu, vous verrez apparaître une interface puis cliquez sur « images » dans l'interface principale. Cliquez sur « Bibliothèque » et sélectionnez l es photos que vous souhaitez supprimer. Cliquez sur « Su pprimer » pour terminer.	*comment supprimer des photos de mon iphone*	**"Connect your iPhone using the USB cable. As soon as your device has been recognized, you will see an interface appear then click on" images "in the main interface. Click on" Library "and select the photos you wish to delete. Click on "Delete" to finish."**	**how to delete photos from my iphone**
Il faut en moyenne 9 à 12 mois pour construire une maison. Il est généralement admis qu'il faut entre 9 à 12 mois pour construire une maison.	*combien de temps pour construire une maison*	**It takes an average of 9 to 12 months to build a house. It is generally accepted that it takes between 9 to 12 months to build a house.**	**how long to build a house**

Table 11 mBART Predictions in French

Model	Passage	Question
mBART	*En la Contabilidad Financiera, el denominado deterioro contable es un gasto previsible, no sistemático y reversible en toda empre sa. Las normas internacionales de Contabilidad establecen que los activos deben someterse a la prueba del deterioro para reflejar con mayor fidelidad su valor en los estados contables.* ***<Es>***	***qu'est ce que le deterioro contable?***
xPN	*En la Contabilidad Financiera, el denominado "deterioro contable" es un gasto previsible, no sistemático y reversible en toda empre sa. Las normas internacionales de Contabilidad establecen que los activos deben someterse a la prueba del "deterioro" para reflejar con mayor fidelidad su valor en los estados contables.*	***que es el deterioro contable?***

Table 12 Zero-Shot Generation Comparison between mBART and xProphetNet

Passage	Question	Translated passage	Translated question
Minecraft bietet ab sofort Cross-Platform-Gameplay. Seit dem jüngsten Update ist die Windows 10 Edition kompatibel mit der Pocket Edition und den Versionen für Xbox-Konsolen. Maximal sind fünf Mitspieler möglich. Alle drei beziehungsweise vier (Xbox 360 und Xbox One) Plattformen funktionieren aber offenbar nicht untereinander.	*ist Minecraft kompatibel mit xbox 360*	**Minecraft now offers cross-platform gameplay. Since the recent update, Windows 10 Edition is compatible with Pocket Edition and versions for Xbox consoles. A maximum of five players are possible. However, all three or four (Xbox 360 and Xbox One) platforms apparently do not work with each other.**	**Is Minecraft compatible with xbox 360**
Sumac, il più potente antiossidante al mondo. Il Sumac è una spezia orientale dal sapore acidulo e ricchissima di antiossidanti. La spezia mediorientale che si trova anche in Sicilia. Il suo vero nome (arabo) è summāq e si tratta di una spezia davvero super antiossidante – la Rhus Coriaria – originaria del Medioriente.	*che è il più potente antiossidante al mondo*	**Sumac, the most powerful antioxidant in the world. Sumac is an oriental spice with a sour taste and rich in antioxidants. The Middle Eastern spice is also found in Sicily. His real name (Arabic) it is summ-q and it is a really super antioxidant spice – the Rhus Coriaria – native to the Middle East.**	**which is the most powerful antioxidant in the world**
Публичная кадастровая карта (ПКК) является актуальным рабочим инструментом и материалом для риэлторов, судебных инстанций, муницип альных органов власти, частных лиц, страхователей, банковских и прочих финансовых учреждений, юристов и т.д. Официальным держателем и регистратором базы данных на публичной кадастровой карте России является официальный	*является публичной кадастровой картой*	**Public cadastral map (PCC) is a relevant working tool and material for realtors, courts, municipality authorities, individuals, insurers, banking and other financial institutions, lawyers, etc. and the database registrar on Russia's public cadastral map is the official**	**is a public cadastral map**

Table 13 Zero-Shot Generation Examples

VII. References

A. C. Graesser, P. C. (2005). Autotutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

Alexander M. Rush, S. C. (2015). A neural attention model for abstractive sentence summarization. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal*, 379-389.

Bahdanau, D. C. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. . *CoRR, abs/1409.0473.*

Bao, H. D. (2020). UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. . *ArXiv, abs/2002.12804.*

Cho, K. M. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. . *EMNLP*.

Clark, J. C. (2020). TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. . *ArXiv, abs/2003.05002.*

Conneau, A. K. (2020). Unsupervised Cross-lingual Representation Learning at Scale. . *ArXiv, abs/1911.02116.*

Devlin, J. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT.*

Dong, L. Y. (2019). Unified Language Model Pre-training for Natural Language Understanding and Generation. *NeurIPS.*

Du, X. S. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. *ArXiv, abs/1705.00106.*

Gu, J. L. (2016). Incorporating Copying Mechanism in Sequence-to-Sequence Learning. *ArXiv, abs/1603.06393.*

Gülçehre, Ç. A. (2016). Pointing the Unknown Words. . *ArXiv, abs/1603.08148.*

Heilman, M. &. (2009). Question Generation via Overgenerating Transformations and Ranking. *Technical Report CMU-LTI-09-013, Language Technologies Institute, Carnegie Mellon University*.

Hinton, G. V. (2015). Distilling the Knowledge in a Neural Network. *ArXiv, abs/1503.02531*.

Hochreiter, S. &. (1997). Long Short-Term Memory. . *Neural Computation, 9*, 1735-1780.

Hosking, T. &. (2019). Evaluating Rewards for Question Generation Models. *ArXiv, abs/1902.11049.*

Huang, H. L. (2019). Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. . *EMNLP/IJCNLP*.

Kelvin Xu, J. B. (2015). Show, attend and tell: Neural image caption generation with visual attention. *ICML volume 14*, 77-81.

Kim, Y. L. (2019). Improving Neural Question Generation using Answer Separation. *AAAI.*

Kingma, D. &. (2015). Adam: A Method for Stochastic Optimization. . *CoRR, abs/1412.6980.*

Lample, G. &. (2019). Cross-lingual Language Model Pretraining. *ArXiv, abs/1901.07291*.

Li, J. G. (2019). Improving Question Generation With to the Point Context. *ArXiv, abs/1910.06036.*

Liang, Y. D. (2020). XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. . *ArXiv, abs/2004.01401*.

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. . *ACL*.

Liu, Y. G. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. . *ArXiv, abs/2001.08210.*

M. A. Walker, O. R. (2001). Spot:a trainable sentence planner. *NAACL*.

Papineni, K. R. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL*.

Peters, M. N. (2018). Deep contextualized word representations. . *ArXiv, abs/1802.05365.*

Radford, A. W. (2019). Language Models are Unsupervised Multitask Learners.

Raffel, C. S. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. . *ArXiv, abs/1910.10683.*

Rennie, S. M. (2017). Self-Critical Sequence Training for Image Captioning. . *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* , 1179-1195.

Ross, J. R. (1967). Constraints on variables in syntax. *Massachusetts Institute of Technology. Dept. of Modern Languages and Linguistics. Thesis. 1967.*

Rus, V. &. (2010). The First Question Generation Shared Task Evaluation Challenge. . *In Proceedings of the 6th International Natural Language Generation Conference (INLG '10). Association for Computational Linguistics, USA*, 251-257.

Sennrich, R. H. (2016). Neural Machine Translation of Rare Words with Subword Units. . *ArXiv, abs/1508.07909.*

Song, L. W. (2018). Leveraging Context Information for Natural Question Generation. *NAACL-HLT*.

Srinivasan Iyer, I. K. (2016). Summarizing source code using a neural attention model. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers). Association for Computational Linguistics, Berlin, Germany*, 2073-2083.

Srivastava, N. H. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res., 15*, , 1929-1958.

Sutskever, I. V. (2014). Sequence to Sequence Learning with Neural Networks. *ArXiv, abs/1409.3215*.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 415-433.

Thang Luong, H. P. (2015a). Effective approaches to attention-based neural machine translation. *EMNLP*, 1412-1421.

Vanderwende, L. (2008). The Importance of Being Important: Question Generation. *Proceedings of the 1st Workshot on the Question Generation Shared Task Evaluation Challenge*.

Vaswani, A. S. (2017). Attention is All you Need. *ArXiv, abs/1706.03762*.

Wolfe, J. (1976). Automatic question generation from text - an aid to independent study. *SIGCSE*.

Zhang, S. &. (2019). Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering. . *ArXiv, abs/1909.06356*.