Questions Generation from Relations

By: Gil Levy, Liad Levi-Raz

Submitted as final project report for the NLP course

Reichman University, 2022

1 Introduction

Motivated by the paper "Supervised Relation Classification as Two-way Span-Prediction" (https://arxiv.org/pdf/2010.04829.pdf) we propose a method to generate questions that represent different relations in a sentence. This paper proposes an improved method for relation classification: The main idea is introducing questions (one/few per relation) to a question / answer model and classifying the relation based on the questions that lead to the correct answer. In case of many questions per relation and tie score between number of correct answers any heuristics can be applied to classify the relation. The question generation is an important building block of this architecture .A question represents a relation in a form that the answer is either the subject entity or the object entity.

1.1 Related Works

The paper denoted above proposes 2 ways for two questions per relation. The paper uses a simple template for the question generation task. Following a few template examples:

Relation Name	Question 1	Question 2
date_of_birth	When was e1 born?	Who was born in e2
parents	Who are the parents of e1	Who are the children of e2
cause_of_death	How did e1 died	How died by e2

The benefits of using automatic question generation are: 1. Avoids the need to define a new template for new relations. 2. Generate more diverse questions. This may improve the relation classification task as proposed by the paper.

2 Solution

2.1 General approach

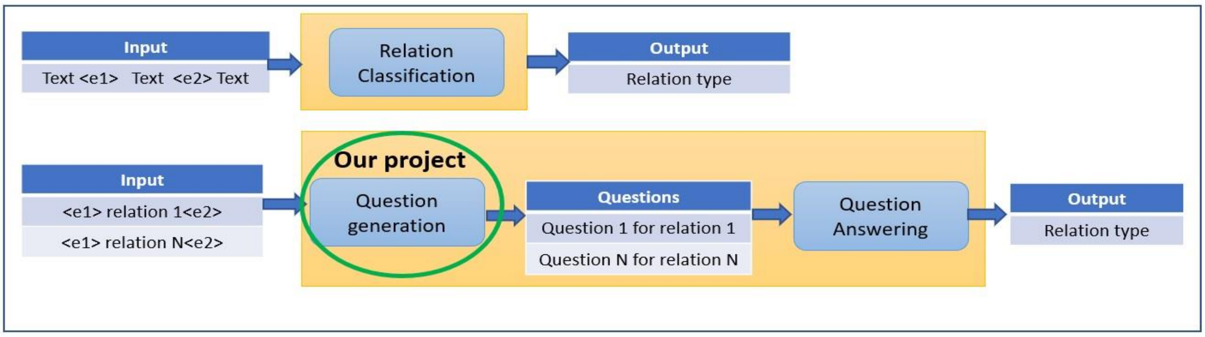


Figure1 Relation classification architectures

Our project proposes an automatic tool for the question generation block. It can be used for a relation classification task as described above as well as for other NLP tasks requiring a relation search. We have developed the solution in two phases:

1. **Basic Model**: Question generator model where the input is the triple of {Subject entity , Relation type , Object entity} and the output is a question representing this relation for the input objects.

2. **Enhanced Model**: A model that generates additional diverse questions by either playing with the generation parameters (rephrasing) or using additional external data related to the Subject and / or object entities from a knowledge database

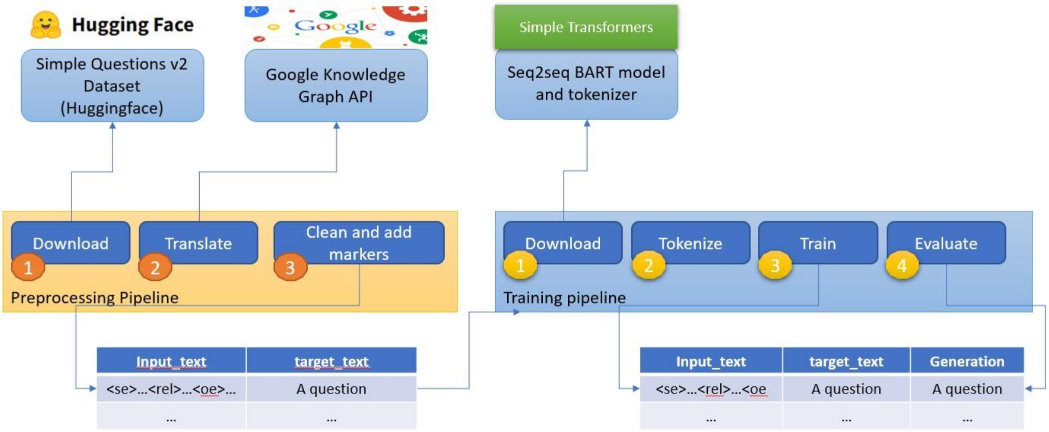


Figure2 - The Basic Model

(Numbers indicate the order of operations)

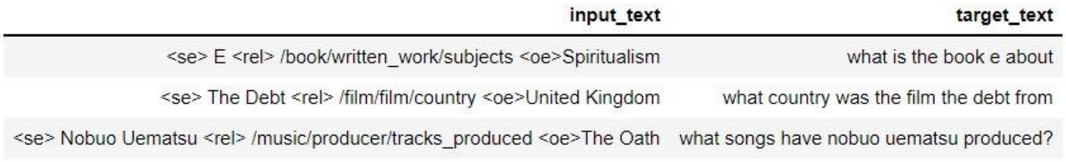
2.2 Design

2.2.1 Dataset and Preprocessing

The dataset used is Simple Questions v2 dataset from Huggingface https://huggingface.co/datasets/simple_questions_v2 . It contains IDs of 2 entities, a relation between them and a question. The entities and relations are encoded using the legacy 'Freebase' database IDs (the Freebase database APIs is no longer available, however some of the information can be retrieved using the Google Knowledge Graph API)

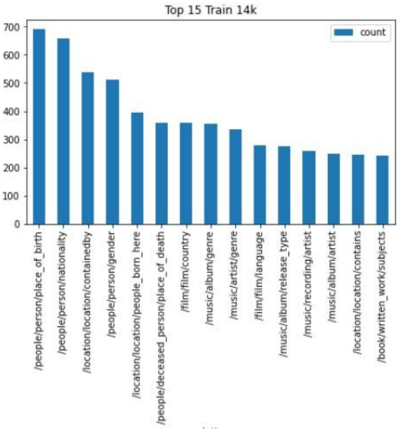
The preprocessing steps includes the following:

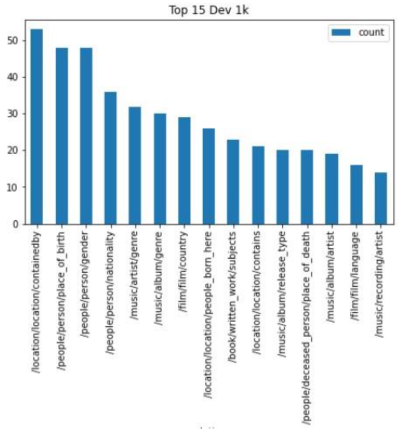
- Translating the Freebase IDs using Google Knowledge Graph API and replacing the 'Subject entity' and the 'Object entity', the Freebase 'Relation' IDs are kept as is
- Adding three special markers <se>,<rel> and <oe> for separation between the objects and the relation
  - Produce three dataframes Train (14K) , Dev (1K) and Test (2K) with the below structure :

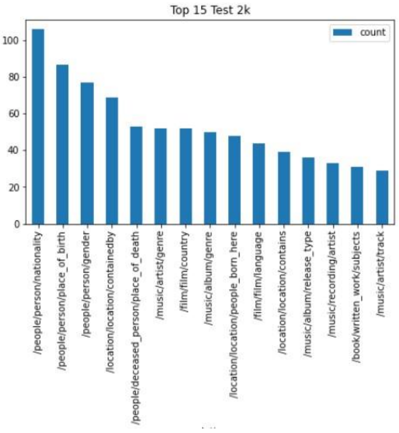


input_text	target_text
<se> E <rel> /book/written_work/subjects <oe>Spiritualism	what is the book e about
<se> The Debt <rel> /film/film/country <oe>United Kingdom	what country was the film the debt from
<se> Nobuo Uematsu <rel> /music/producer/tracks_produced <oe>The Oath	what songs have nobuo uematsu produced?

The following figures show the most common relations in the train and test data sets.







Note that there are many relations that have only a few/single instances in the train. Moreover, in the test there are relations that **do not exist in the train** (for example there is only a single instance in the dataset for this specific relation). This issue is further discussed in the evaluation section later.

Scale: The size of the original dataset is 108K . Approximately ~70% of the object IDs are found in the translation preprocessing stage , a total of ~70K entries - which is more than enough for our needs and our computation power capabilities. We finally sampled **14K , 1K , 2K** for train , dev , test to meet our computational resources.

2.2.2 Model

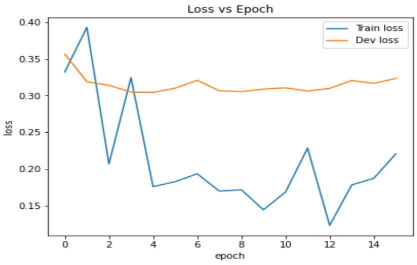
We have used a seq2seq model based on the BART. BART model is optimized for tasks involving NLP generation. BART does not follow the convention of using a similar structure to BERT which mainly consists of a stack of encoder blocks. Instead, BART uses the standard Transformer architecture (i.e. both encoder and decoder blocks are involved). The pretrained model we have selected is "facebook/bart-large" . With BartTokenizer we have checked the **size of the train inputs** (after tokenization) and **limited the model input size** to 99.5% percentile + 5. By this approach we have removed some length outliers and significantly reduced the model size and training run time.

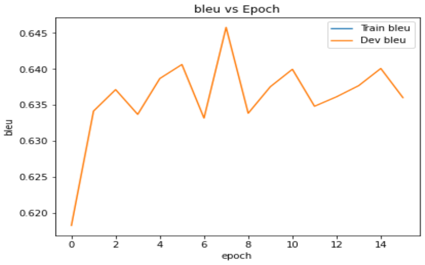
Note: while looking for the "best" pretrained model we also experimented with other pretrained models such as: "mbart","facebook/bart-large-xsum","facebook/bart-large-cnn","facebook/mbart-large-50-one-to-manymmt","facebook/mbart-large-cc25",”facebook/mbart-large-50", but eventually **"facebook/bart-large"** was the selected. We also trained a model with T5 architecture instead of BART in which and got similar BLEU score results, and similar quality of generated questions

Text generation: There are various options to control the text generation: Greedy search , beam search , sampling (with temperature) , top k sampling and top p sampling. For the **basic model,** where we are only looking for a single result "question", we have found the following set of parameters yielding the best questions. : num_beams= 4 , max_length= 20 , length_penalty= 1.0 , repetition_penalty= 1.0 , num_return_sequences= 1. Note that we have observed minor to negligible differences for some of the parameters. Generating diverse questions is further discussed in the "generation questions".

2.2.3 Training

We have performed training with the following hyper parameters: LR = 1-e5 , Batch size = 8 (limited by our compute recourse) , Early stopping. During training we have evaluated the results by bleu score on both train and validation datasets. We have enabled learning for the entire layers of the model. This option yields better results (usually the training takes about 1 hour (+) and 15 to 20 epochs, while the loss keeps improving on both train and dev, the average BLEU score on the Dev set is not improving)





Note: we achieved similar results both by using the “simpletransformers” library and by training our own BART model from scratch (our own trainer and Bart tokenizer). Therefore, we have used “simpletransformers” for the entire project for simplicity. The model reached average **BLEU scores of 0.7 / 0.64** for the training / test datasets

3 Experimental results

3.1 Basic Model

In this section we have analyzed the results by two categories. 1. The BLEU score that is widely used for seq2seq tasks 2. By “Eye examination”. As observed by a few examples provided below the BLEU score may not always capture the quality of the generated question. The bleu score works well in one direction. That is a high score relates to high quality questions however a low score may not always relate to low quality generated questions.

3.1.1 Examples

Following are a few examples of questions generated by the model vs the original label

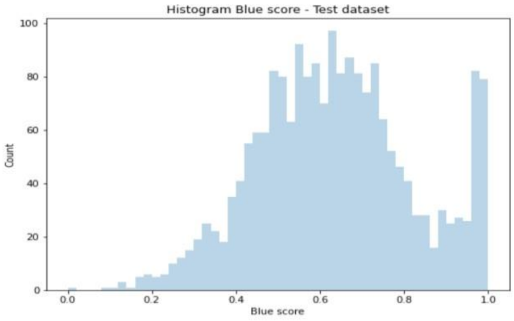
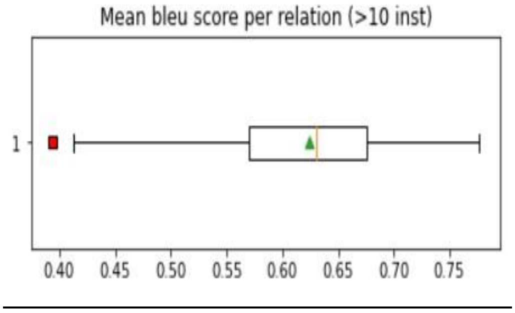
No	Label question	Predicted question
1	Name a notable teacher	Name a teacher.
2	where was debi gliori born	where was debi gliori born
3	What type of music is watch out! classified as	what kind of music is watch out!
4	where in australia was james plimsoll's place of death?	where did james plimsoll die
5	what is the capital of chemnitzer land	what is the capital of chemnitzer land

Following a few examples of questions generated by the model for relations that we invented

No	Relation	Predicted question	Comment
1	<se> Tel Aviv <rel> /location/location/contains <oe>Yarkon Stream",	what is a stream in tel aviv (Known relation)	Known relation
2	The lecturer of the natural language course in Reichman university is Amir"	what university is amir a professor of natural language	
3	<se> Liad_Gil <rel> authors <oe> NLP project	what is the name of the invention by liad gil	

3.1.2 BLEU score analysis

Blue Score Histogram (Test)	Mean BLEU score distribution per relation (TEST)

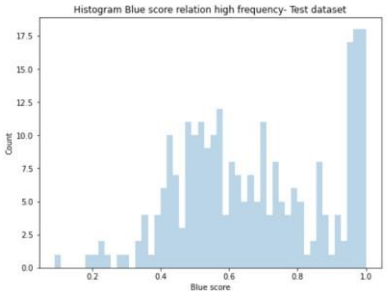
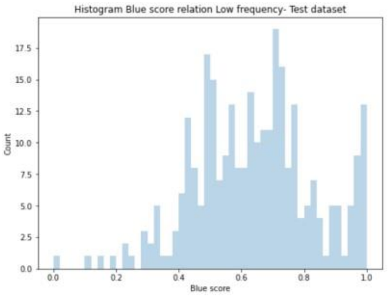
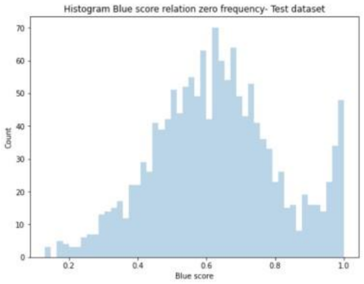
Note: As the bleu score of relations are not highly correlated to their frequency in the train (See analysis below) we can analyze the mean bleu per relation in the test.

Conclusions:

- The blue score has a wide distribution , especially many with high score close to 1
- The mean distribution of the bleu score per relation is not that wide. That is the relation itself has moderate impact on the bleu score

Bleu score as function of relation frequency in train

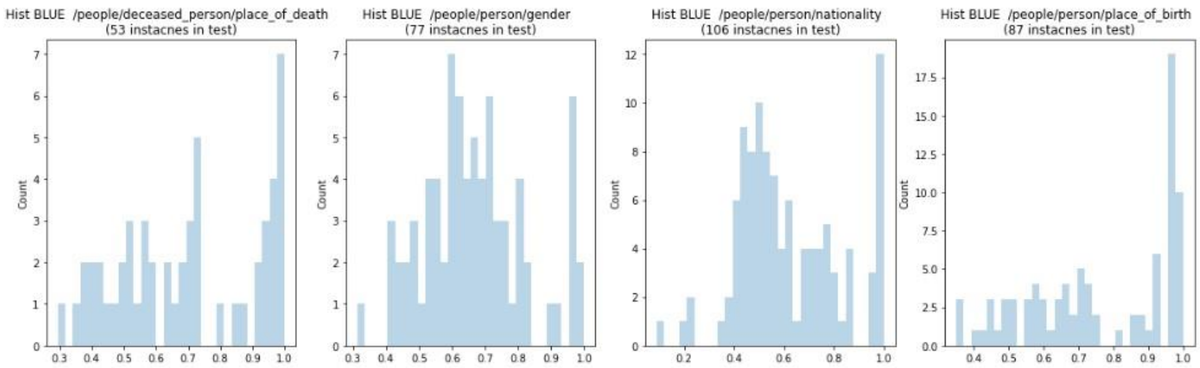
While the mean bleu for the test is **0.63,** we can see a wide distribution of the score over the test instances. To analyze the effect of the relation frequency in train on the test results we check the bleu score for high / low / no relation frequency in the train:

Highest frequent relations	**Lowest frequent relations**	**Relations in Test only**
		
13% of test instances. These test instances relate to the 3 highest relation frequency in the **Train** which are 1887 / 14000 of the train	15% of test instances. These instances relate to the 636 lowest relation frequency in the **TRAIN** dataset which are 2212/ 14000 of the train	1447 of test instances. These test instances have **no** relation in the **TRAIN** dataset
BLEU mean: 0.673	BLEU mean: 0.64	BLEU mean: 0.63

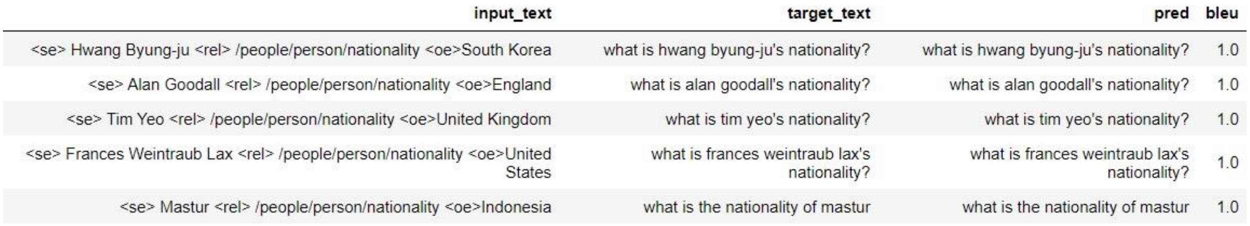
Conclusion: The model generates **high quality questions** for relations in the test that are **rare** / do not appear in the train dataset

Bleu score distribution for high frequency relations in the **test**

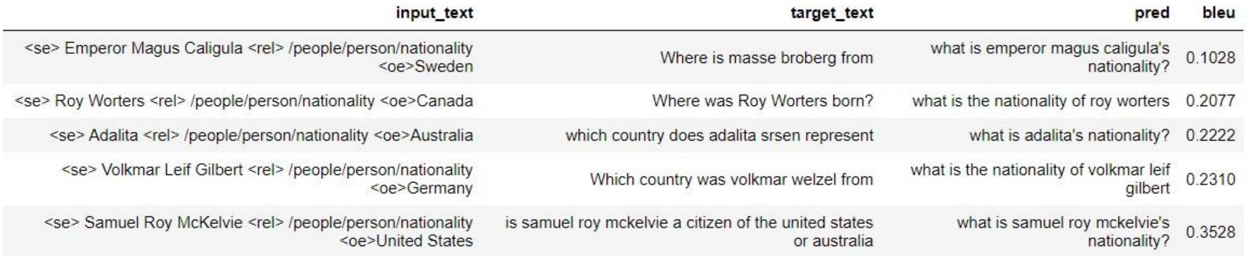
For relations that have high frequency in the test we analyze the bleu distribution. If distribution is wide, we can check how to utilize the case of high score instances for our needs



Below are examples for high and low bleu scores for one of the above relations. The bleu score is low (except for the first example) as the model tends to generate the same questions based on templates while the labels are more diverse. Generating diverse questions is further discussed in the Enhance model



input_text	target_text	pred	bleu
<se> Hwang Byung-ju <rel> /people/person/nationality <oe>South Korea	what is hwang byung-ju's nationality?	what is hwang byung-ju's nationality?	1.0
<se> Alan Goodall <rel> /people/person/nationality <oe>England	what is alan goodall's nationality?	what is alan goodall's nationality?	1.0
<se> Tim Yeo <rel> /people/person/nationality <oe>United Kingdom	what is tim yeo's nationality?	what is tim yeo's nationality?	1.0
<se> Frances Weintraub Lax <rel> /people/person/nationality <oe>United States	what is frances weintraub lax's nationality?	what is frances weintraub lax's nationality?	1.0
<se> Mastur <rel> /people/person/nationality <oe>Indonesia	what is the nationality of mastur	what is the nationality of mastur	1.0

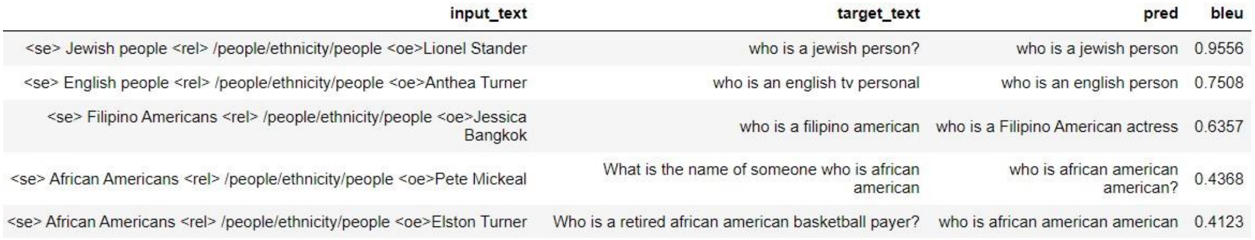


input_text	target_text	pred	bleu
<se> Emperor Magus Caligula <rel> /people/person/nationality <oe>Sweden	Where is masse broberg from	what is emperor magus caligula's nationality?	0.1028
<se> Roy Worters <rel> /people/person/nationality <oe>Canada	Where was Roy Worters born?	what is the nationality of roy worters	0.2077
<se> Adalita <rel> /people/person/nationality <oe>Australia	which country does adalita srsen represent	what is adalita's nationality?	0.2222
<se> Volkmar Leif Gilbert <rel> /people/person/nationality <oe>Germany	Which country was volkmar welzel from	what is the nationality of volkmar leif gilbert	0.2310
<se> Samuel Roy McKelvie <rel> /people/person/nationality <oe>United States	is samuel roy mckelvie a citizen of the united states or australia	what is samuel roy mckelvie's nationality?	0.3528

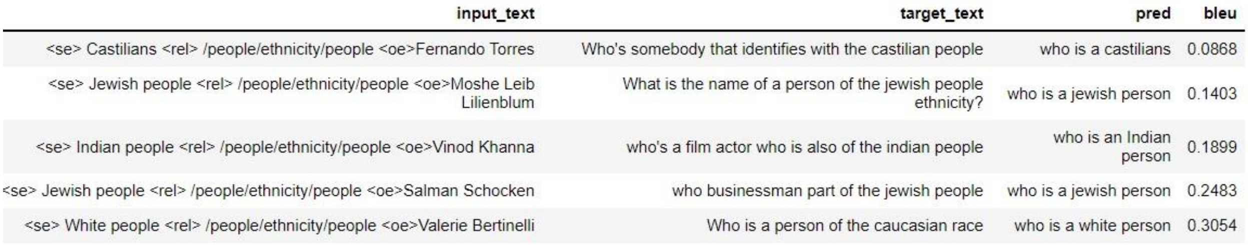
Bleu score vs “By Eye” score (More examples in the NB)

Input	Target	Prediction	Bleu core
<se> Eswatini <rel> /location/location/partially_contains <oe>Emlembe	which places partially contains swaziland?	what is a town in eswatini	0.0001
<se> Rebel <rel> /book/written_work/author <oe>Tom Hayden	rebel was written by what author	who wrote rebel	0.1261
<se> Jewish people <rel> /people/ethnicity/people <oe>Moshe Leib Lilienblum	What is the name of a person of the jewish people ethnicity?	who is a jewish person	0.1403
<se> Castilians <rel> /people/ethnicity/people <oe>Fernando Torres	Who's somebody that identifies with the castilian people	who is a castilians	0.0868

Further look at the relation: ‘/people/ethnicity/people‘ that yields low quality questions also by eye, shows that for some questions (the last 3) the labels have data that are not part of the input text. This might be improved by adding more information to the input text that describe the objects



input_text	target_text	pred	bleu
<se> Jewish people <rel> /people/ethnicity/people <oe>Lionel Stander	who is a jewish person?	who is a jewish person	0.9556
<se> English people <rel> /people/ethnicity/people <oe>Anthea Turner	who is an english tv personal	who is an english person	0.7508
<se> Filipino Americans <rel> /people/ethnicity/people <oe>Jessica Bangkok	who is a filipino american	who is a Filipino American actress	0.6357
<se> African Americans <rel> /people/ethnicity/people <oe>Pete Mickeal	What is the name of someone who is african american	who is african american american?	0.4368
<se> African Americans <rel> /people/ethnicity/people <oe>Elston Turner	Who is a retired african american basketball payer?	who is african american american	0.4123

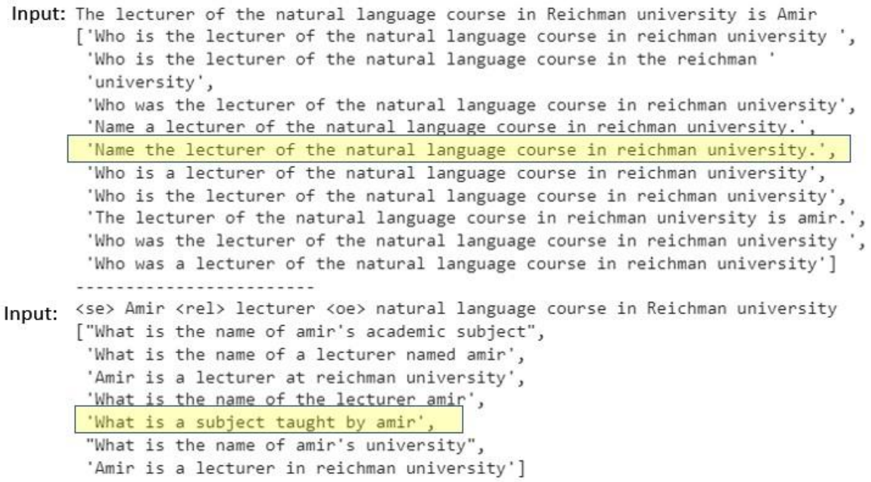


input_text	target_text	pred	bleu
<se> Castilians <rel> /people/ethnicity/people <oe>Fernando Torres	Who's somebody that identifies with the castilian people	who is a castilians	0.0868
<se> Jewish people <rel> /people/ethnicity/people <oe>Moshe Leib Lilienblum	What is the name of a person of the jewish people ethnicity?	who is a jewish person	0.1403
<se> Indian people <rel> /people/ethnicity/people <oe>Vinod Khanna	who's a film actor who is also of the indian people	who is an Indian person	0.1899
<se> Jewish people <rel> /people/ethnicity/people <oe>Salman Schocken	who businessman part of the jewish people	who is a jewish person	0.2483
<se> White people <rel> /people/ethnicity/people <oe>Valerie Bertinelli	Who is a person of the caucasian race	who is a white person	0.3054

Note: For high bleu score the quality of the questions is always good (Details in the NB)

3.2 Enhanced Model

We experimented various options of the model generation: 'top_k' , 'top_p' and 'beam_search'. Eventually we kept the 'beam_search' option with the following parameters:num_beams= 4, max_length= 22,min_length= 4, length_penalty= 2.0, early_stopping= True,repetition_penalty= 2.0,num_return_sequences= 1), but the other options didn't yield significantly different results for our dataset (they were quite good as well). One observation is that when increasing the num_beams and num_return_sequences we observed **a nice paraphrasing capability** of the model, to generate rephrased questions out of the box (questions with the same meaning), for example:

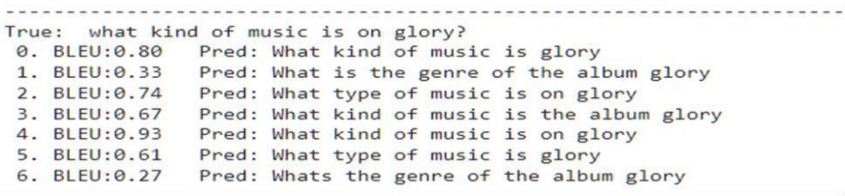


```
Input: The lecturer of the natural language course in Reichman university is Amir
       ['Who is the lecturer of the natural language course in reichman university ',
        'Who is the lecturer of the natural language course in the reichman '
        'university',
        'Who was the lecturer of the natural language course in reichman university',
        'Name a lecturer of the natural language course in reichman university.',
        'Name the lecturer of the natural language course in reichman university.',
        'Who is a lecturer of the natural language course in reichman university',
        'Who is the lecturer of the natural language course in reichman university',
        'The lecturer of the natural language course in reichman university is amir.',
        'Who was the lecturer of the natural language course in reichman university ',
        'Who was a lecturer of the natural language course in reichman university']
       -----------------------
Input: <se> Amir <rel> lecturer <oe> natural language course in Reichman university
       ["What is the name of amir's academic subject",
        'What is the name of a lecturer named amir',
        'Amir is a lecturer at reichman university',
        'What is the name of the lecturer amir',
        'What is a subject taught by amir',
        "What is the name of amir's university",
        'Amir is a lecturer in reichman university']
```

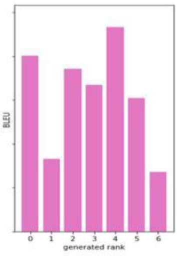
(additional examples in the notebooks)

One conclusion that we learned, is that **beam search** works well here, probably because the **length** of the desired generation is more or less **predictable** (in a text 2 text task like ours)

In addition we wanted to analyze the relation between the **rank** of a generated question and its bleu score (**1st rank** is the first generated question returned, then 2nd,3rd,...) , basically we expected to see that the first returned sequence has a higher BLEU score than the other, however this is not always the case , when **we do rephrasing -** we care less about the BLEU score, because we want to achieve **diversity**, here is a good example where the 2nd question is in a high quality but with a low BLEU score:



```
----------------------------------------------------------------------
True:  what kind of music is on glory?
 0. BLEU:0.80    Pred: What kind of music is glory
 1. BLEU:0.33    Pred: What is the genre of the album glory
 2. BLEU:0.74    Pred: What type of music is on glory
 3. BLEU:0.67    Pred: What kind of music is the album glory
 4. BLEU:0.93    Pred: What kind of music is on glory
 5. BLEU:0.61    Pred: What type of music is glory
 6. BLEU:0.27    Pred: Whats the genre of the album glory
```



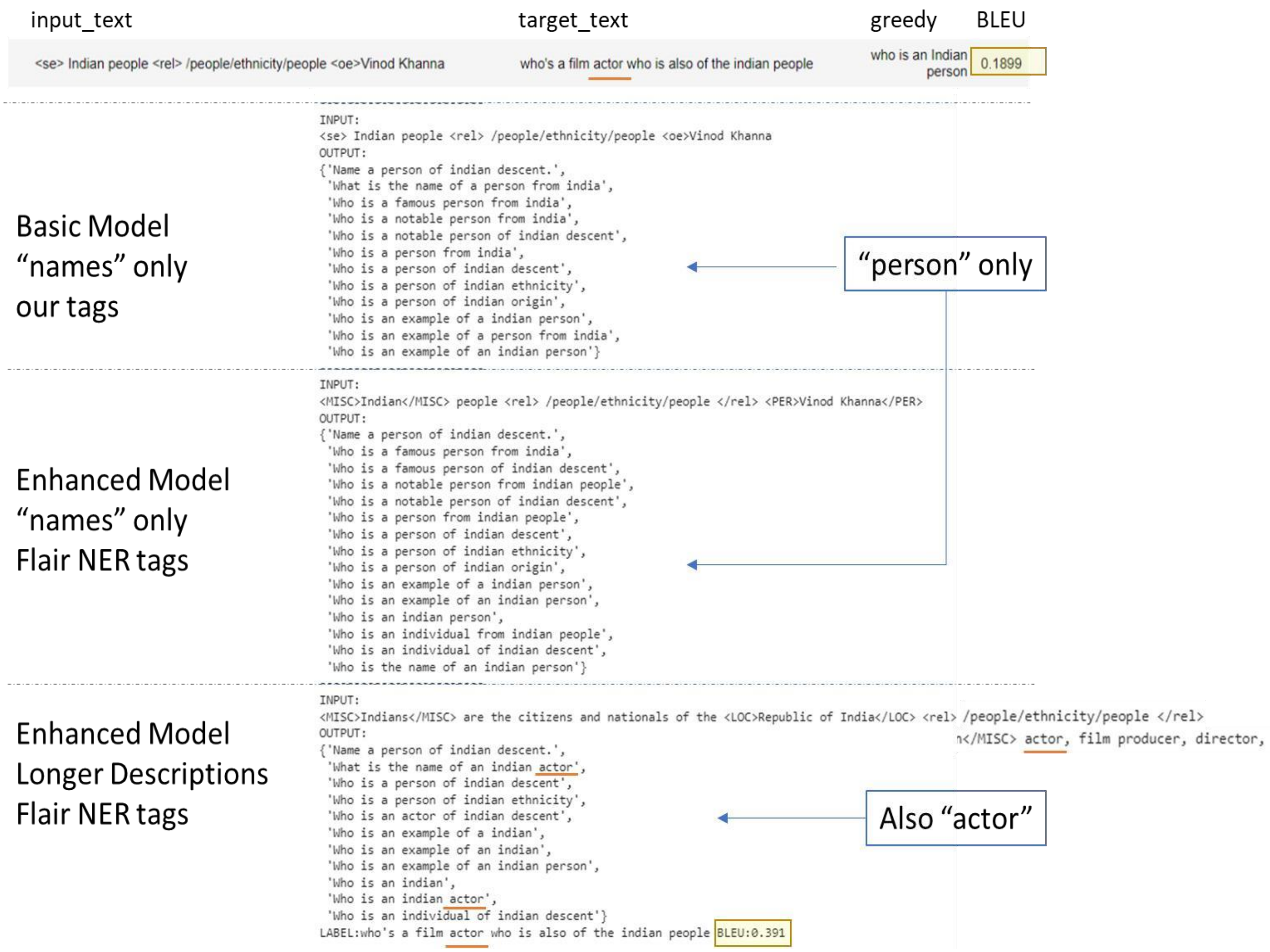
To generate more interesting questions, we enriched our dataset in the following way:

- For every **Subject Entity** or **Object Entity** in our translated datasets, we queried the **Google Knowledge Graph API** again, and retrieved a **longer description** of the entity

- We selected **a sentence or two** from the description, to get a **shorter texts -** we hoped that the additional text will contain important words that **also appear in the label**
- We run the text through the **FlairLNP NER** model to tag the entity descriptions (process took a few hours…)

Eventually we got the following new dataset, this time with the new tags ('LOC','MISC','ORG','PER') added the tokenizer and run the same Basic BART model again:

We compared the Basic model generation with our simple tagging, and the enhanced version with the Flair NERs and longer descriptions, we were able to see **more diversity in the rephrasing** and in cases where the **additional text also appeared in the label** also improved quality (and BLEU). for example here is a Beam Search generation comparison (look at the word "actor"):



input_text	target_text	greedy	BLEU
<se> Indian people <rel> /people/ethnicity/people <oe>Vinod Khanna	who's a film actor who is also of the indian people	who is an Indian person	0.1899

```
INPUT:
<se> Indian people <rel> /people/ethnicity/people <oe>Vinod Khanna
OUTPUT:
{'Name a person of indian descent.',
 'What is the name of a person from india',
 'Who is a famous person from india',
 'Who is a notable person from india',
 'Who is a notable person of indian descent',
 'Who is a person from india',
 'Who is a person of indian descent',
 'Who is a person of indian ethnicity',
 'Who is a person of indian origin',
 'Who is an example of a indian person',
 'Who is an example of a person from india',
 'Who is an example of an indian person'}
```

```
INPUT:
<MISC>Indian</MISC> people <rel> /people/ethnicity/people </rel> <PER>Vinod Khanna</PER>
OUTPUT:
{'Name a person of indian descent.',
 'Who is a famous person from india',
 'Who is a famous person of indian descent',
 'Who is a notable person from indian people',
 'Who is a notable person of indian descent',
 'Who is a person from indian people',
 'Who is a person of indian descent',
 'Who is a person of indian ethnicity',
 'Who is a person of indian origin',
 'Who is an example of a indian person',
 'Who is an example of an indian person',
 'Who is an indian person',
 'Who is an individual from indian people',
 'Who is an individual of indian descent',
 'Who is the name of an indian person'}
```

```
INPUT:
<MISC>Indians</MISC> are the citizens and nationals of the <LOC>Republic of India</LOC> <rel> /people/ethnicity/people </rel>
OUTPUT:                                                                              n</MISC> actor, film producer, director,
{'Name a person of indian descent.',
 'What is the name of an indian actor',
 'Who is a person of indian descent',
 'Who is a person of indian ethnicity',
 'Who is an actor of indian descent',
 'Who is an example of a indian',
 'Who is an example of an indian',
 'Who is an example of an indian person',
 'Who is an indian',
 'Who is an indian actor',
 'Who is an individual of indian descent'}
LABEL:who's a film actor who is also of the indian people BLEU:0.391
```

4 Discussion

By training a BART seq2seq model we have succeeded to generate automatically questions representing relations. The model can generate high quality questions for relations that do not appear in the train set. There is no major difference in the quality of the questions among different relations. To generate question with diversity, we have controlled the generation options of the seq2seq model. To further generate interesting questions, we have added more information to the inputs and retrained the model

5 Code

Original dataset Simple Questions v2	Preprocessed datasets - For Basic model - For enhanced model	Preprocessing notebook	Basic Model notebook	Enhanced Model notebook (same as Basic Model but trained with enhanced dataset)

References

- https://huggingface.co/blog/how-to-generate - explains the generate options nicely
- https://tungmphung.com/a-review-of-pre-trained-language-models-from-bert-roberta-to-electradeberta-bigbird-and-more/ - Review of differ Bert models
- https://simpletransformers.ai/ - Simple transformers library