
# M915 NLU & NLG

(Fall Semester 2024)

## Assignment 2: Question Answering systems

Report

Kleopatra Karapanagiotou (lt12200010)

# Contents

## 1 Introduction/Task Description

In this assignment four Question Answering systems were developed for the SQuAD v.1 dataset with the use of Zero-, One- and Few-shot learning schemes on the Flan-T5-small Language Model. The first version of SQuAD is a reading comprehension dataset, and it is designed for extractive question answering, where the answer to each question is extracted by a span in the given context.

## 2 Data processing and analysis

### 2.1 Pre-processing

The dataset was loaded directly from HugginFace and no preprocessing was implemented.

### 2.2 Analysis

For the task of Question Answering it was interesting to inspect the categories of questions posed by the crowdworkers by adopting an initial superficial classification approach, based on the pronoun that each Question starts. This of course was a naive approach, due to the high lexical and syntactic variation in the posed questions and the high diversity of answer types reported in the initial paper (Rajpurkar P., et.al. 2016), it was nevertheless a starting point for a wider view of the dataset content.

In the following table we see the distribution of the questions of the train dataset in the inspected question types. It is expected that the category 'Other' contains the second highest percentage of instances in the train set, since this category concentrates the whole linguistic variance of the human posed questions, especially since the crowd workers were asked to pose questions with their own words and not copy-pasting content of the given passage.

| | |
|---|---|
| Other | 18234 |
| What | 35475 |
| When | 5056 |
| How | 7689 |
| Where | 3083 |
| Which | 4052 |
| Who | 7714 |
| Why | 1109 |
| Be/Do | 601 |
| Whose | 184 |
| Whom | 22 |

```
Questions for class 'Other':
1. From whom did the Scythians of Central Asia discover Hellenistic culture?
2. The Cambrian period was how long ago?
3. Over what scandal did the Chinese government lose in public opinion?
4. It is also common to connect an airport and a city with what?
5. Nasser led which country in 1961?
6. France was now in a desperate position, what did they do?
7. In what year did executives at Trump Entertainment Group say they were considering selling the Taj?
8. After the fall of Napoleon, many countries retained what system of law?
9. Residence that speak German as their mother tongue and  families have been in place for generations are often consider?
10. To which reporter did Kanye West express regret for his remark about President Bush?
```

### 2.3 Data partitioning in train/test/validation

The train dataset was splitted, so that the first 95% was used for choosing demonstration examples, and the remaining 5% was used as validation set for finetuning purposes in choosing the number of shots for the few-shot setting. For inference the actual validation set was used, since the test set of SQuAD is not publicly available.

| train | val | test |
|---|---|---|
| 83219 | 4380 | 10570 |

## 3 LLM choice and Experiments

Flan-T5 inherits its name from the T5 encoder-decoder architecture and the Flan instruction- tuning method used on a collection of data sources with a variety of instruction template types. It is an open-source LLM pre-trained on a variety of language tasks, including question answering and due to its instruction-tuning it can perform various zero-shot NLP tasks, as well as few-shot in-context learning tasks with the goal of learning mappings between sequences of text, i.e., text-to-text. Since the SQuAD v1 dataset was not a dataset that the model has seen before during pretraining or instruction-finetuning, according to the listed datasets[1], it was considered an appropriate fit for our zero-, one- and few-shot experiments. Due to computational capacity the small version was used in all experiments.

### 3.1 Experiments

#### 3.1.1 Zero-Shot

In the zero-shot setting no instruction was given to the model, rather only the posed question after providing the relevant text passage. In the picture below we see a first correct attempt of zero-shot learning. The only provided configuration was the maximum number of new tokens to generate (max_new_tokens=25)

```
Input:  ["In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to F
-----------------------------------------------------------------------------------------------
Human question:  ['In what year was Father Edward Sorin given two years to create a college?']
-----------------------------------------------------------------------------------------------
Model Output:  1842
```

#### 3.1.2 One-Shot

In the one-shot setting the prompt template includes

- a keyword followed by the text passage (of the train_set) in the new line
- new line
- an instruction followed by the posed question in the new line
- new line
- a key-phrase followed by the true answer (including the ‘answer_start’ information) in the new line.
- new line
- a keyword followed by a new text passage (of the val/test set) in the new line.
- new line
- an instruction followed by the posed question in the new line
- a key-phrase denoting the expectation for an extracted answer based on context and the single demonstration above.

```
Passage:
Notre Dame is known for its competitive admissions, with the incoming class enrolling in fall 2015 admitting 3,577 from a pool of 18,156 (19.7%)

Read the passage above and answer the following question:
What percentage of students were admitted to Notre Dame in fall 2015?

The answer is:
{'text': ['19.7%'], 'answer_start': [138]}

Passage:
In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to Father Edward Sorin of the Congregation of the Holy Cros

Read the passage above and answer the following question:
In what year was Father Edward Sorin given two years to create a college?

The answer is:
```

[1] https://arxiv.org/pdf/2210.11416.pdf

### 3.1.3 Few-Shot

The same prompt template is applied in the few-shot setting, with the addition of multiple demonstrations. Below an example of the 3-hot prompt template:

```
Passage:
Notre Dame is known for its competitive admissions, with the incoming class enrolling in fall

Read the passage above and answer the following question:
What percentage of students were admitted to Notre Dame in fall 2015?

The answer is:
{'text': ['19.7%'], 'answer_start': [138]}

Passage:
About 80% of undergraduates and 20% of graduate students live on campus. The majority of the g

Read the passage above and answer the following question:
How many dorms for males are on the Notre Dame campus?

The answer is:
{'text': ['15'], 'answer_start': [350]}

Passage:
The Rev. Theodore Hesburgh, C.S.C., (1917–2015) served as president for 35 years (1952–87) of

Read the passage above and answer the following question:
What was the size of the Notre Dame endowment when Theodore Hesburgh became president?

The answer is:
{'text': ['$9 million'], 'answer_start': [262]}

Passage:
In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to Father

Read the passage above and answer the following question:
In what year was Father Edward Sorin given two years to create a college?

The answer is:
```

In our experiment, 9 randomly chosen shots were used for inference and the details of this choice are reported in the hyperparameter tuning table of trials below.

#### 3.1.3.1 Hyper-parameter tuning

| | **3-shot** | **6-shot** | **9-shot** |
|---|---|---|---|
| **EM** | 72.99 | 72.37 | **73.97** |
| **F1** | 85.85 | 85.36 | **86.41** |

It seems that from 3 to 6 demonstrations there is no significant difference in performance. With 9 demonstrations we get slightly better results. This leaves us with the idea of experimenting with even higher numbers of demonstrations, to see if this slight improvement increases.

#### 3.1.3.2 Extra Experiment

The question type classification reported above in the 'Analysis' section was further employed to create a new dataset containing one example of each question type. Inspecting the new_dataset content below, we can report various answer types captured by each question: Named Entities (Person, Location, Organization), complex and simple Noun Phrases, Dates, Boolean answers, and adverbs. This classification managed to cover a good percentage of the reported question types in the original SQuAD paper (Rajpurkar P., et.al. 2016). The experiment was conducted with the same prompt template. No hyperparameter tuning was done, the answers were generated only for the test set.

```
Context: Architecturally, the school has a Catholic character. Atop the Main Buil
Question: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
Answers: {'text': ['Saint Bernadette Soubirous'], 'answer_start': [515]}
--------------------------------------------------
Context: Architecturally, the school has a Catholic character. Atop the Main Buil
Question: What is in front of the Notre Dame Main Building?
Answers: {'text': ['a copper statue of Christ'], 'answer_start': [188]}
--------------------------------------------------
Context: As at most other universities, Notre Dame's students run a number of new
Question: When did the Scholastic Magazine of Notre dame begin publishing?
Answers: {'text': ['September 1876'], 'answer_start': [248]}
--------------------------------------------------
Context: As at most other universities, Notre Dame's students run a number of new
Question: How often is Notre Dame's the Juggler published?
Answers: {'text': ['twice'], 'answer_start': [441]}
--------------------------------------------------
Context: The university is the major seat of the Congregation of Holy Cross (albe
Question: Where is the headquarters of the Congregation of the Holy Cross?
Answers: {'text': ['Rome'], 'answer_start': [119]}
--------------------------------------------------
Context: The university is the major seat of the Congregation of Holy Cross (albe
Question: Which prize did Frederick Buechner create?
Answers: {'text': ['Buechner Prize for Preaching'], 'answer_start': [675]}
--------------------------------------------------
Context: As of 2012[update] research continued in many fields. The university pre
Question: Who was the president of Notre Dame in 2012?
Answers: {'text': ['John Jenkins'], 'answer_start': [80]}
--------------------------------------------------
Context: In August, the couple attended the 2011 MTV Video Music Awards, at which
Question: Why was the broadcast the most-watched in history?
Answers: {'text': ['Her appearance'], 'answer_start': [417]}
--------------------------------------------------
Context: Montana's personal income tax contains 7 brackets, with rates ranging fr
Question: Does Montana have a sales tax?
Answers: {'text': ['no'], 'answer_start': [113]}
--------------------------------------------------
Context: All of Chopin's compositions include the piano. Most are for solo piano,
Question: Whose music did Frédéric admire the most and thus provide influence on
Answers: {'text': ['J. S. Bach, Mozart and Schubert'], 'answer_start': [620]}
--------------------------------------------------
Context: Chopin's relations with Sand were soured in 1846 by problems involving h
Question: Whom did Sand's daughter Solange become engaged to?
Answers: {'text': ['Auguste Clésinger'], 'answer_start': [149]}
--------------------------------------------------
```

### 3.4 Evaluation

For evaluation the official SQUAD v.1.1 evaluation script[2] was used and contains the following functions:

- **normalize_answer(s):** which converts text to lowercase, removes punctuation, articles, and extra whitespace, so the text in references and predictions has the same form.
- **f1_score (prediction, ground_truth):** measures the average overlap between prediction and ground truth answer. The prediction and ground truth are treated as bags of tokens, and their F1 is computed**.**
- **exact_match_score(prediction, ground_truth)**: returns a boolean value (T/F) indicating whether the normalized version of the prediction matches the normalized version of the ground truth exactly.
- **metric_max_over_ground_truths(metric_fn, prediction, ground_truths):** iterates over each ground truth answer in the given dataset, calculates a metric (EM/F1) for each prediction and ground truth, saves the scores in a list and returns the maximum metric score over all ground truth answers for each question in the given dataset

[2] https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py

- **evaluate (gold_answers, predictions)**: iterates over each ground truth and predicted answer, computes the maximum Exact Match and F1 over all of the ground truth answers for a given question, and then averages over all of the questions to get the final EM and F1 percentage.

### 3.4.1 Exact Match| F1-score

| | **0-shot** | **1-shot** | **9-shot** | **Exp-11-shot** |
|---|---|---|---|---|
| **EM** | 73.03 | 81.40 | 80.93 | **82.34** |
| **F1** | 81.87 | 88.28 | 88.87 | **89.84** |

### 3.4.2 Comparison of Results with the SQuAD1.1 dev Benchmark[3]

Our learning schemes with Flan-T5 Small gave results comparable to the following models trained on bigger portions of the train set compared to our few shot experiments and it was worth showing a comparison.

| | **Model** | **EM** | **F1** |
|---|---|---|---|
| Zero-shot | QANet (Yu et al., 2018) | 73.6 | 82.7 |
| One shot/9-shot | BERT-base(single) (Devlin et al.,2019) | 80.8 | 88.5 |
| 11-shot | BERT-Large-uncased-PruneOFA (Zafrir et al., 2021) | 83.3 | 90.2 |

## 4 Results and Overall Analysis

From the evaluation results above we can reach several conclusions regarding each experiment setting:

- LLM performance increased significantly with the addition of a single task demonstration through a prompt template in both evaluation metrics, which means that the model learned more patterns with One-shot learning.
- The addition of multiple demonstrations, which ranged from 3 to 9, did not indicate significant increase, neither during finetuning nor during testing. Especially the equal results between 1-shot and 9-shot may be a good indicator for experimenting with bigger window sizes between the shots (e.g. 10-20-30). I chose to experiment with a small range due to the unknown inference time[4] required and the lack of computational resources for all experiments.

### 4.1 Best trial

The experiment with 11 demonstrations from the classified questions gave indeed comparably higher scores in both metrics. This improvement is though small to be able to explain, if it was due to higher variation in the quality of the input or simply due to the increase in demonstrations. Considering again the small window between the demonstrations in the finetuning and taking into consideration few shot learning practices that recommend the K-value to be up to "the maximum amount allowed by the model's context window (2048, which fits 10 to 100 examples)" (Brown et al., 2020), it would be

[3] https://paperswithcode.com/sota/question-answering-on-squad11-dev

[4] Inference on 9-shot and 11-shot experiment took more than 2.30 hours each.

interesting to experiment with k-values biggen than 10 (e.g.10,20,30-shot learning) to see if performance increases. Another point for further experimentation is the use of more prompt templates, starting from some examples found in Github[5]. Finally, given that "few-shot improves by model size" (Brown et al., 2020), we can experiment with larger versions of Flan T5, as long as our computational power allows us to do so.

## References


- Abdollahpour, M. M. (2022, September 9). *Zero-shot Question Answering with Large Language Models in Python*. NLPlanet. https://medium.com/nlplanet/zero-shot-question-answering-with-large-language-models-in-python-9964c55c3b38
- Balci, P. (2023, August 21). *LLM — Few-Shot Learning*. Medium. https://medium.com/@balci.pelin/llm-few-shot-learning-d7df1d2c4446
- Balci, P. (2023, October 19). *LLM — Inference*. Medium. https://medium.com/@balci.pelin/llm-inference-222c8e8a6ba7
- Balcı, P. (2024, January 31). *pelinbalci/LLM_Notebooks*. GitHub. https://github.com/pelinbalci/LLM_Notebooks/tree/main
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Hesse, C. (2020). *Language Models are Few-Shot Learners*. https://arxiv.org/pdf/2005.14165.pdf
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/pdf/1810.04805.pdf
- *Question answering - Hugging Face NLP Course*. (n.d.). Huggingface.co. Retrieved February 15, 2024, from https://huggingface.co/learn/nlp-course/chapter7/7?fw=pt
- Otten, N. V. (2023, June 29). *Few-shot Learning Explained & Step-by-step How To Python Tutorial*. Spot Intelligence. https://spotintelligence.com/2023/06/29/few-shot-learning/
- Le Scao, T., & Rush, A. (2021). *How Many Data Points is a Prompt Worth?* (pp. 2627–2636). https://aclanthology.org/2021.naacl-main.208.pdf
- Tian, E., & Huang, K. (n.d.). *Prompting for Few-shot Learning Motivation and Related Work*. Retrieved February 15, 2024, from https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec05.pdf
- *T5*. (n.d.). Huggingface.co. https://huggingface.co/docs/transformers/model_doc/t5
- *Zero-shot prompting for the Flan-T5 foundation model in Amazon SageMaker JumpStart | AWS Machine Learning Blog*. (2023, April 3). Aws.amazon.com. https://aws.amazon.com/blogs/machine-learning/zero-shot-prompting-for-the-flan-t5-foundation-model-in-amazon-sagemaker-jumpstart/
- Yu, A., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., & Le Google Brain, Q. (n.d.). *Published as a conference paper at ICLR 2018 QANET: COMBINING LOCAL CONVOLUTION WITH GLOBAL SELF-ATTENTION FOR READING COMPRE- HENSION*. https://arxiv.org/pdf/1804.09541.pdf


[5] https://github.com/aws/amazon-sagemaker-examples/blob/main/introduction_to_amazon_algorithms/jumpstart-foundation-models/text2text-generation-flan-t5.ipynb