Open in app   Get started

tds   Published in Towards Data Science · Follow

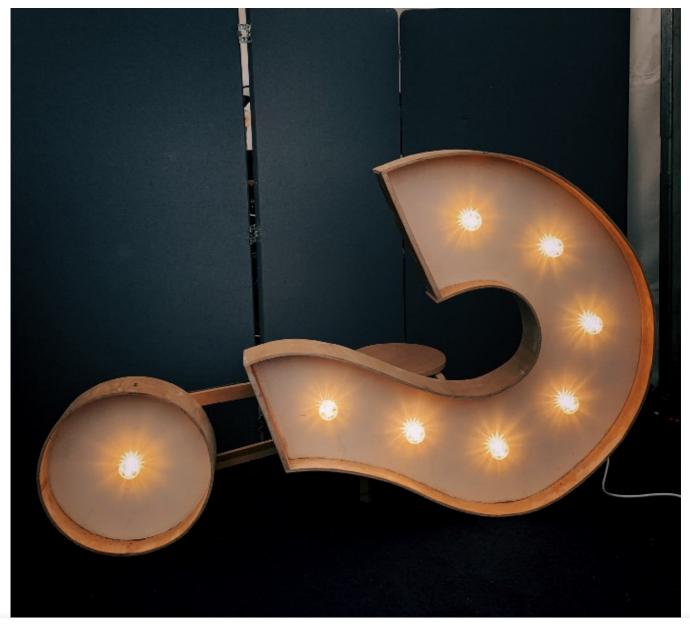You have **1** free member-only story left this month. Sign up for Medium and get an extra one

Susumu Okazawa, PhD · Follow

Feb 24, 2021 · 6 min read ★

Swedish Question Answering with BERT

How to fine-tune non-English BERT

Open in app   Get started

Question answering is one of the most well-studied tasks in the field of NLP (Natural Language Processing), and a highlight of the last few years has been the emergence of BERT.

BERT and other Transformer-based models are now available for a variety of languages. However, for relatively minor languages, e.g. Swedish, the available models are limited or do not exist at all. We faced several challenges when trying to fine-tune Swedish BERT to a question answering task. The purpose of this post is to document how we solved them.

The main focus of this post will be on the question answering task using the SQuAD 2.0.

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. SQuAD2.0 [Pranav Rajpurkar, Robin Jia & Percy Liang, 2018] combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions.

Three ways of constructing Swedish QA systems with BERT-family

If you want to use BERT-family to do a question answering task in Swedish (or your preferred non-English language), you can come up with three ways.

1. Translate Swedish questions into English, process it with English BERT, and translate English answers back into Swedish
   👍There are many resources available in English, including fine-tuned models.
   👎Translation is performed for every inference, which increases the computational cost. There is a possibility of mistranslation.

2. Process Swedish questions with multilingual models
   👍Multilingual models have also been the subject of much research, so there are many resources available.
   👎Multilingual models have a larger model size than monolingual models, which makes them inefficient when you only want to process Swedish.

👍Relatively small models can be obtained for efficient processing.
👎Mistranslation of the dataset may lead to unintended bias.

In this article, we will focus on the third option, which is to use Swedish SQuAD to build a Swedish QA BERT.

Translation of SQuAD 2.0

We have translated SQuAD 2.0 automatically using the Google Translate API. It sounds like straightforward task but it is not the case for the following reasons.

- The span that determines the start and end of the answer in the context may change after translation.

- If the context and the answer are translated independently, the translated answer may not be included in the translated context.

Let me explain the first point in more detail. For example, the *Oxygen* text in SQuAD 2.0 dev has the following context (excerpted from a whole sentence), question, and answer.

> **Context**
> Oxygen is a chemical element with symbol O and atomic number 8.
>
> **Question**
> The atomic number of the periodic table for oxygen?
>
> **Answer**
> 8
>
> - excerpt from SQuAD 2.0 dev, *Oxygen*

In this case, the span of the answer is from 60 to 61, which indicates the position of the start and end of the answer in the excerpted context.

This context could be translated to Swedish as follows.

⌂　　　　　　　　　　　　🔍　　　　　　　　　　　　✎

Now, the correct span of the answer is from 53 to 54. Thus, **not only do we need to translate the context and the answer, we also need to correctly track their position in the context.**

Regarding the second point, for instance, we have the following context, question, and answer in the *Oxygen* text.

> **Context**
> One of the first known experiments on the relationship between combustion and air was conducted by the 2nd century BCE Greek writer on mechanics, Philo of Byzantium.
>
> **Question**
> In what year was the first known experiments on combustion and air conducted?
>
> **Answer**
> 2nd century BCE
>
> - excerpt from SQuAD 2.0 dev, *Oxygen*

Translating this context by Google, at least for now, yields the following output.

```
Ett av de första kända experimenten om förhållandet mellan
förbränning och luft utfördes av den grekiska författaren om mekanik
under 2000-talet f.Kr. Philo of Byzantium.
```

However, if we translate the answer *2nd century BCE* independently, we get the following.

```
2: a århundradet fvt.
```

Thus, if you translate the context and the answer independently, the ground truth answer may not be found in the context. **We need to translate the context and the answer in a coherent way.**

1. Before the translation, insert the special marker around the answer in the context. For example, the *Oxygen* text in SQuAD 2.0 dev has the following context (excerpted from a whole sentence), question, and answer.

Context

Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning.

Question

Which gas makes up 20.8% of the Earth's atmosphere?

Answer

Diatomic oxygen

- excerpt from <u>SQuAD 2.0 dev</u>, <u>*Oxygen*</u>

In this case, we insert the special marker *[0]* around the answer *Diatomic oxygen*, that yields

```
[0] Diatomic oxygen [0] gas constitutes 20.8% of the Earth's
atmosphere. However, monitoring of atmospheric oxygen levels show a
global downward trend, because of fossil-fuel burning.
```

Note that, the special marker is not limited to *[0]*, but we can use others.

2. Translate the marked context. The result will be like this;

```
[0] Diatomiskt syre [0] gas utgör 20,8% av jordens atmosfär.
Övervakning av syrehalten i atmosfären visar dock en global
nedåtgående trend på grund av förbränning av fossila bränslen.
```

3. Extract the marked sentence from the translated context. It will be the translated answer. The start and end of the marked sentence will be the span of the answer.

⌂   Q   ✎

●❙❙   Open in app   (Get started)

I admit that this strategy is not perfect. Some answers were not found in context, so these inappropriate examples had to be removed. Therefore, the size of the translated dataset is about 90% of the original dataset.

Evaluation on SQuAD 2.0 dev

We fine-tuned Swedish BERT pre-trained by the National Library of Sweden (KB Lab) and evaluated three models on our Swedish translation of SQuAD 2.0 dev dataset.

The first model is the Multilingual XLM-RoBERTa trained on SQuAD by deepset GmbH. This is the example of the second option which is listed above.

The second model is the KB Lab model trained on SQuAD. This model is tagged as "experimental" but it works well.

And the third model is our BERT which is fine-tuned on our Swedish translation of SQuAD 2.0 train dataset.

The results of evaluation are summarized in the below table.

Model	Exact Match	F1
Multilingual XLM-RoBERTa(large)	56.96	**70.78**
Swedish BERT (base, KB Lab)	65.65	68.89
Swedish BERT (base, Ours)	**66.73**	70.11

Our model which has about 110M parameters achieves better scores than the KB Lab model, and establishes a close F1 score compared to the XLM-RoBERTa which has about 550M parameters.

If you are interested in to use the fine-tuned model, the model is available in HuggingFace model hub.

Nobel Prize dataset

We have internally created the Nobel Prize dataset in Swedish for evaluation purposes

🏠   🔍   ✏️

Open in app   Get started

Since this dataset is created independently and the model has not seen this dataset before, this result is realistic if we use the fine-tuned model without further training.

The results of evaluation are summarized in the below table. Our model achieved better results in both exact match and F1 score.

Model	Exact Match	F1
Multilingual XLM-RoBERTa(large)	13.19	60.00
Swedish BERT (base, KB Lab)	32.97	52.41
Swedish BERT (base, Ours)	**46.15**	**61.54**

Summary

A method for properly translating dataset for extractive question answering task was presented. We compared the performance of our fine-tuned model with other fine-tuned models on the translated dataset, and confirmed that our model performs relatively well.

About us

Image by Savantic AB

We are **Savantic AB,** an AI company in Stockholm. We love solving impossible problems!

Open in app   Get started

- https://huggingface.co/KB/bert-base-swedish-cased-squad-experimental

- https://huggingface.co/deepset/xlm-roberta-large-squad2

- https://github.com/susumu2357/SQuAD_v2_sv

- https://huggingface.co/datasets/susumu2357/squad_v2_sv

- https://huggingface.co/susumu2357/bert-base-swedish-squad2

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

Get this newsletter