**Task 1:** **Punctuation and capitalization recovery**

**First finding: The model** [**NeMo**](https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tlt-jarvis/models/punctuationcapitalization_english_bert?msclkid=fb57a262cecf11eca9ff94a43ede49a6) **from Invidia.**

It jointly trains two token-level classifiers on top of a pre-trained language model.

To train it for Catalan we can use a Multilingual BERT corpus.

[Here](https://mccormickml.com/2020/10/05/multilingual-bert/?msclkid=316b47f7cf6e11ecab7477e69431f919) is an example of how a Multilingual BERT works for Arabic language.

The pre-trained model has to be be fine-tuned for Catalan using as much as possible training and evaluation data (text.txt - labels.txt) – as described [here](https://docs.nvidia.com/metropolis/TLT/tlt-user-guide/text/nlp/punctuation_and_capitalization.html).

**Second finding:**

There is an approach to create a [BERT alike model for Catalan language](https://github.com/Softcatala/julibert?msclkid=9b4f30d9cf7111eca674139e46772d8d).

**Third finding:**

[Toward Human-Friendly ASR Systems: Recovering Capitalization and Punctuation for Vietnamese Text](https://www.jstage.jst.go.jp/article/transinf/E104.D/8/E104.D_2020BDP0005/_pdf/-char/en?msclkid=ea17d922cf7911ec9f47c238ba2338a3)

The advantage of this approach it that it doesn’t depend on a pre-trained collection, and it works fast.

The data was “created automatically by crawling through Vietnamese Internet tools…”.

- Thus I assume that the same could be done for Catalan, and the data can be created automatically by crawling through Catalan web sites. Then some further preparation is needed.

However I wasn’t able to go deep and understand it completely, but here are some important points :

„It combines transformer decoder with conditional random field.“

„This combination method aims to improve the ability to extract the longer-range context information so that improves performance while working with the long paragraph.“

„.. presents the superior performance compared to the most recent baseline models in both speed and accuracy.“