Make it clean (P9)

Michele Zenoni (matr. 989482)

Information Retrieval

Prof. Alfio Ferrara

A.A. 2022/2023

Contents

Introduction

Research question and methodology

Experimental results

Concluding remarks

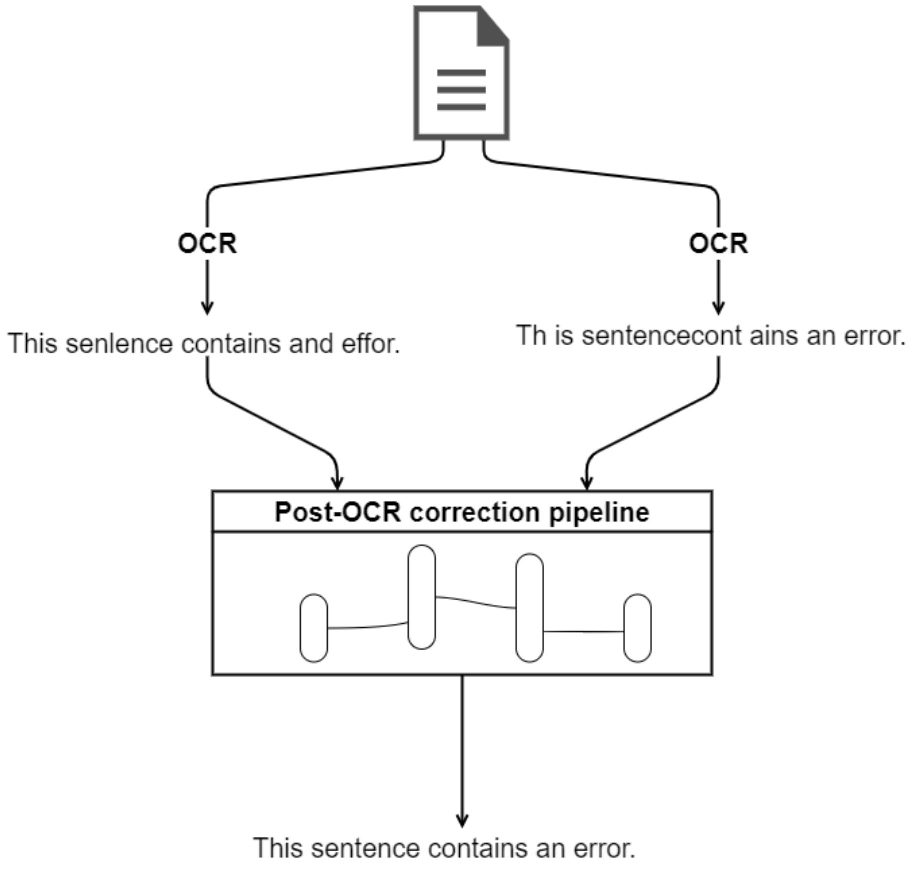
Introduction

Definition, the BART model

Definition

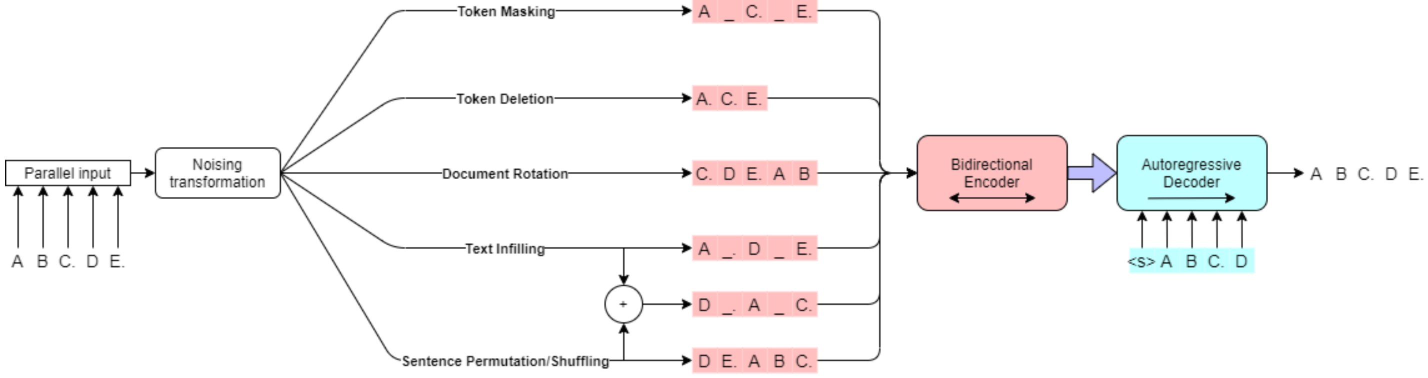
Optical Character Recognition (OCR) is the technique to extract textual information from images (photos or scanned documents) and convert it into machine-encoded text.

It constitutes a major computer science research topic, combining computer vision and **Natural Language Processing (NLP)**.


OCR
OCR
This senlence contains and effor.
Th is sentencecont ains an error.
Post-OCR correction pipeline
This sentence contains an error.

The BART transformer

BART is a **denoising** autoencoder for pre-training sequence-to-sequence models, which can be seen as a generalization of BERT for the use of a **bidirectional encoder** and a GPT with the **left-to-right decoder**.



Research question and methodology

Goals and proposed approach

Purpose of the project

Problem description

The post-OCR correction task could be analyzed as a **spelling correction** problem.

Classical (statistical) approaches don’t take into account the **context** of the sentence.

Goal

Compare different spelling correction strategies to determine if the use of a **transformer architecture** can enhance the statistical approach, improving the overall system quality.

Proposed approach

Implement and compare five spelling correction strategies:

1. Norvig$^+$

Statistical approach enriched with a custom function to deal with simple segmentation errors.

2. SymSpell

Open-source project which provides an optimized statistical spell checker.

3. BART

Fine tuned version of the BART-base model for the spelling correction task, available on the HuggingFace platform.

4. BART + Norvig$^+$

Combination of Norvig$^+$ and BART.

5. BART + SymSpell

Combination of SymSpell and BART (this is expected to be the best strategy).

Experimental results

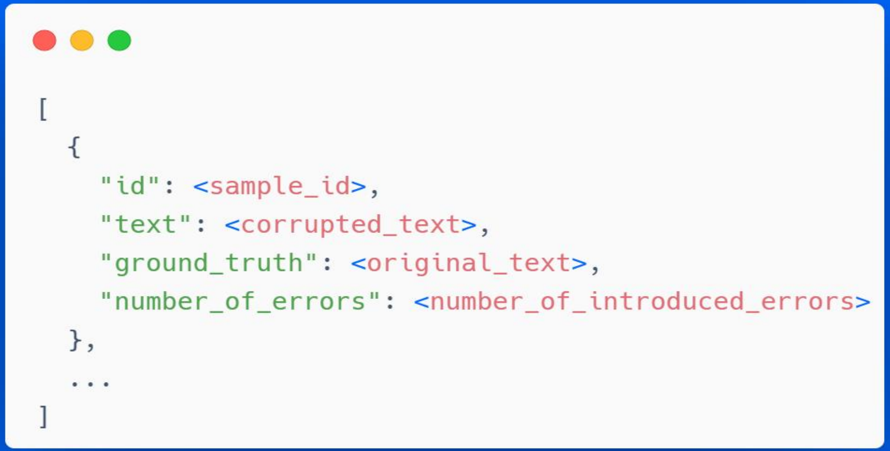
Dataset, metrics and evaluation results

Dataset

Sources & Format

The dataset is **artificially generated** by «corrupting» a ground-truth text.

The original text can be taken using the **Wikipedia API** or sampled from a given **text file** with one sentence per row.

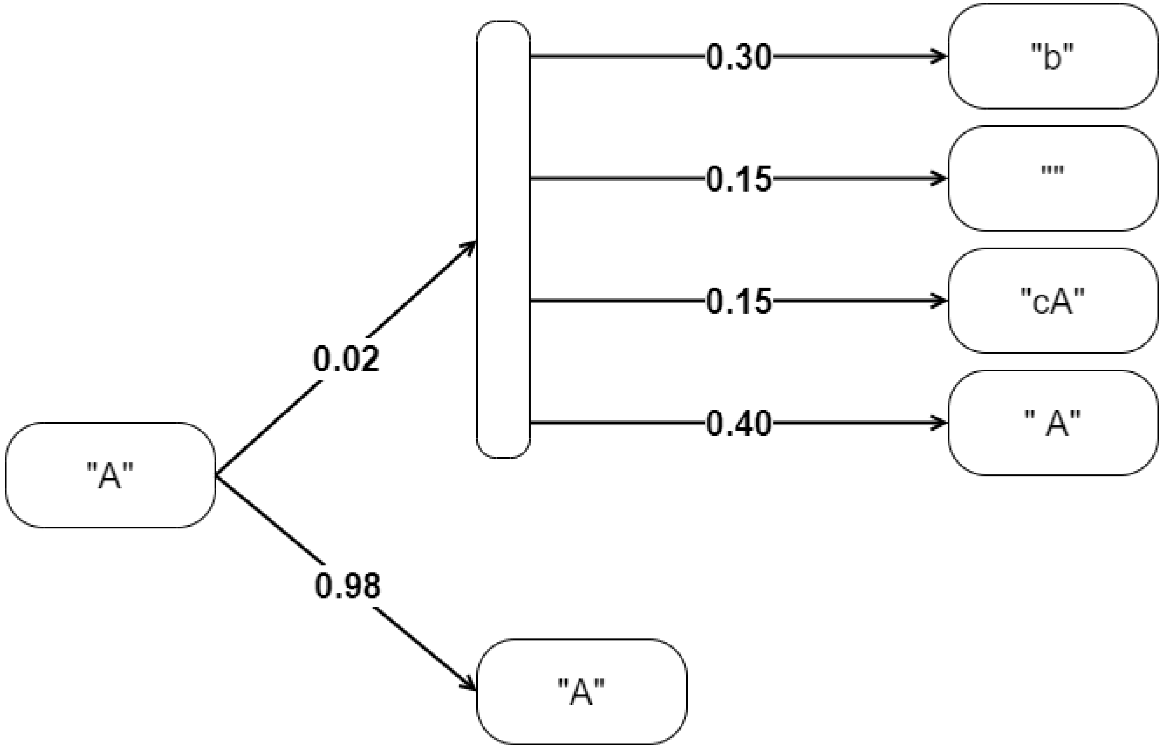


```
[
  {
    "id": <sample_id>,
    "text": <corrupted_text>,
    "ground_truth": <original_text>,
    "number_of_errors": <number_of_introduced_errors>
  },
  ...
]
```

Generation

For each character in the ground truth sentence there's a **2%** probability to apply an editing operation with the following **weights**:

- 30% character substitution
- 15% character deletion
- 15% character insertion
- 40% white space insertion



Evaluation metrics

The performance of the correction strategies has been evaluated according to four metrics:

Evaluation metric	Normalize spaces	Consider punctuation	Case sensitive	Length sensitive
Accuracy	✓	✓	✓	✓
Accuracy (no punct.)	✓	✗	✓	✓
WER	✓	✗	✗	✓
Average CER	✓	✗	✓	✗

The average execution time per sample has also been computed and reported.

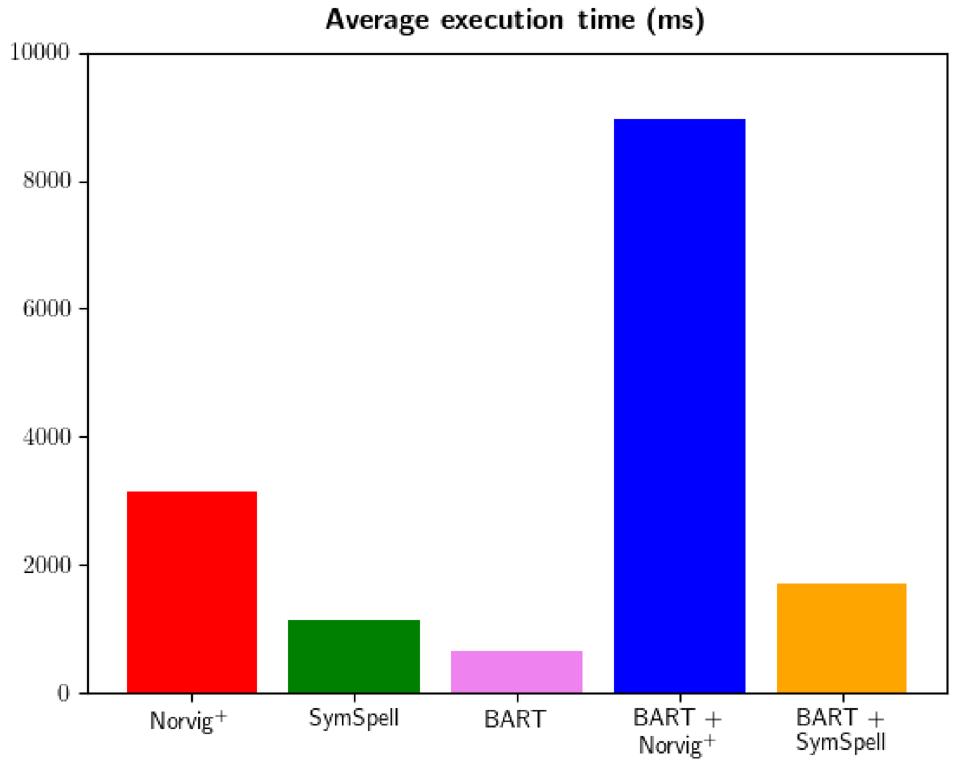
Evaluation results

Correction strategy	Execution time (ms)	Accuracy	Accuracy (no punct.)	WER	Average CER
Norvig$^+$	3156	0.450	0.690	0.404	0.239
SymSpell	1153	0.482	0.739	0.253	0.151
BART	**641**	0.658	0.708	0.401	0.255
BART + Norvig$^+$	8972	0.665	0.777	0.286	0.170
BART + SymSpell	1700	**0.667**	**0.780**	**0.234**	**0.137**

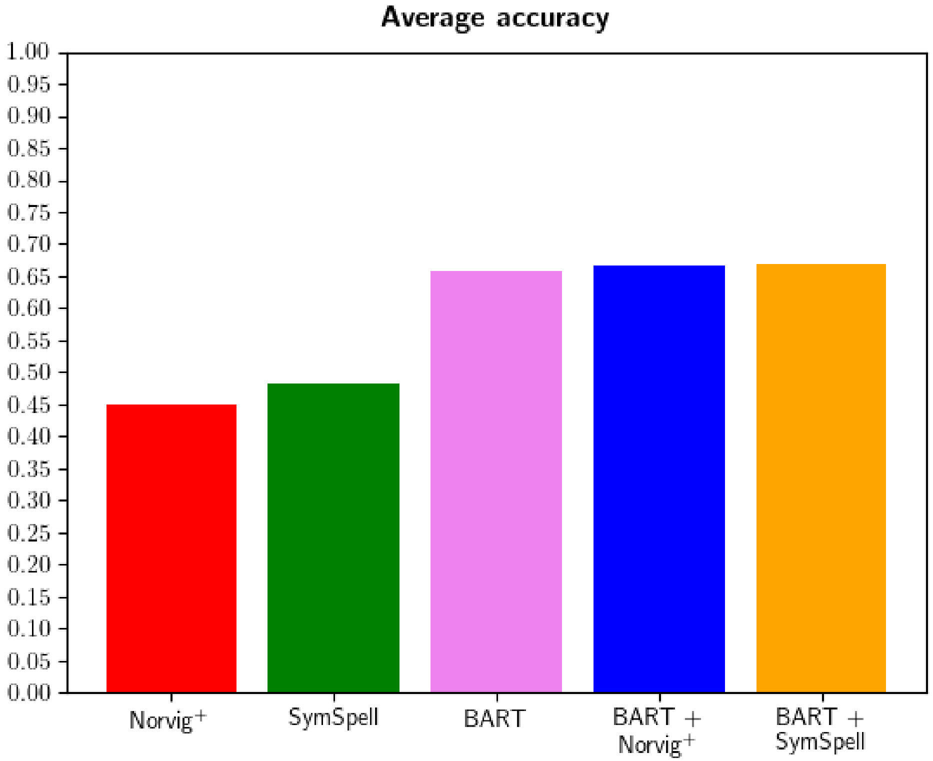
Execution time

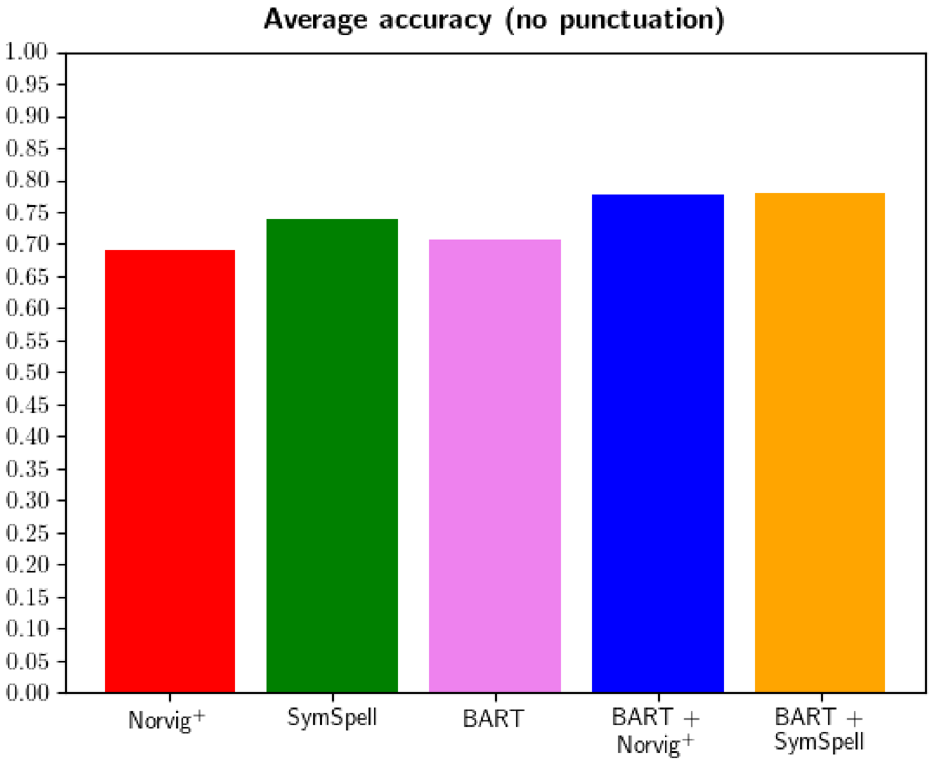
The performance of a spelling correction system should also consider the execution time. The introduction of a deep learning model **does not necessarily cause** a large time overhead.

All evaluations are based on 1000 samples with an average length of 143 characters.

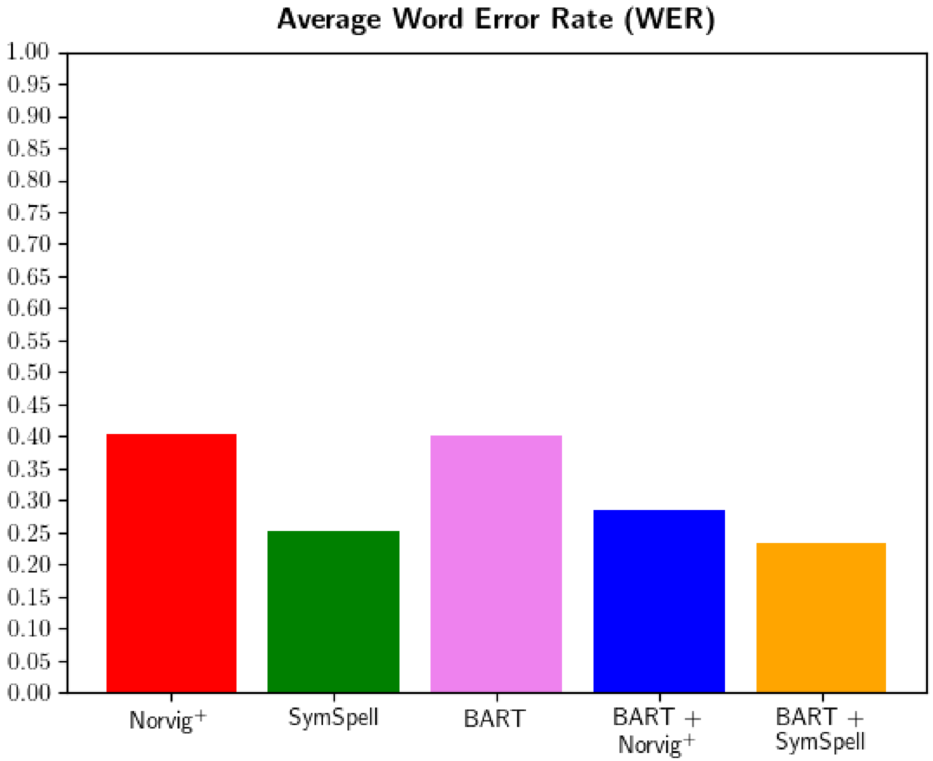


Accuracy

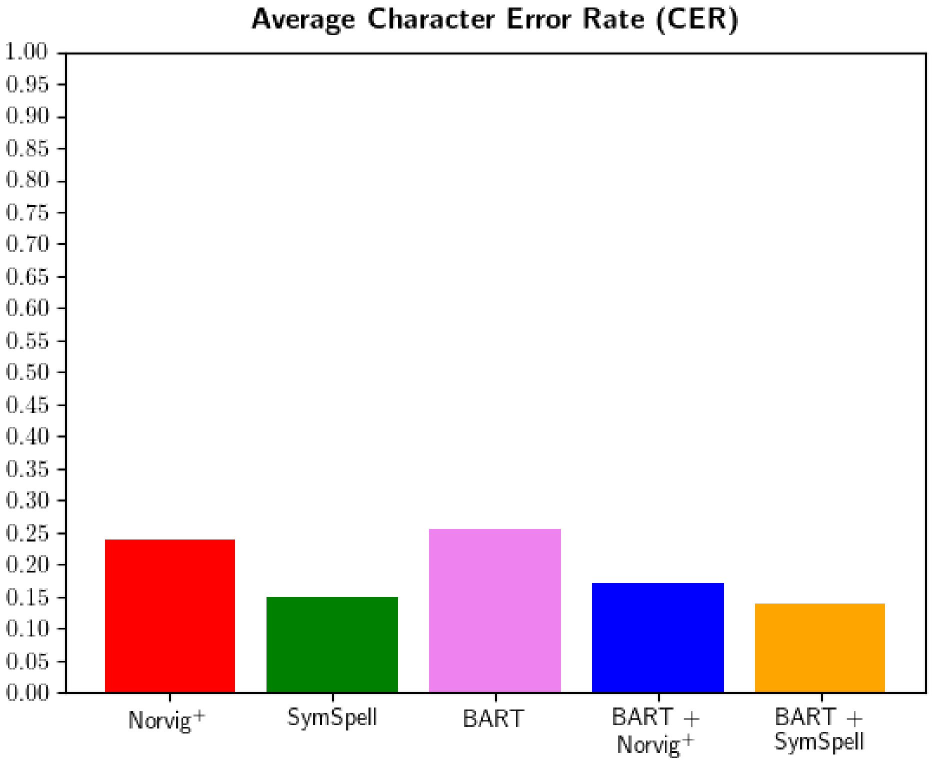

Average accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Norvig+
SymSpell
BART
BART + Norvig+
BART + SymSpell


Average accuracy (no punctuation)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Norvig+
SymSpell
BART
BART + Norvig+
BART + SymSpell

WER & CER



Average Word Error Rate (WER)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Norvig+
SymSpell
BART
BART + Norvig+
BART + SymSpell



Average Character Error Rate (CER)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Norvig+
SymSpell
BART
BART + Norvig+
BART + SymSpell

Concluding remarks

Critical overview, future work

Critical overview

The use of a transformer architecture pre-trained on **denoising** proved to be a valuable support to the classical spelling correction task, particularly when considering complex sentences with **punctuation**.

However, it is necessary to point out some **critical observations**:

- the reached level of accuracy is **not high enough** to make the system been deployed in real world applications
- more tests are needed to find an **optimal configuration** of the system parameters
- the fine-tuned version of BART used for this project has a considerable room for improvement, both in the **training phase** and in the **documentation**

Future work

There are several aspects that could be improved for future work:

- the BART fine-tuning phase should consider a **larger dataset** of artificially corrupted text or real OCRed documents
- the model could be used in combination with more sophisticated statistical techniques for post-OCR correction
- the entire architecture could be **redesigned** to develop a BART model which works at the **character level**

Thank you!

Michele Zenoni

michele.zenoni@studenti.unimi.it

Make-it-clean (github.com)