OCRs for corpus extraction for the Maltese language

Submission for the DocEng 2026 competition.
Contact: Dr Marc Tanti marc.tanti@um.edu.mt
Competition page: https://www.um.edu.mt/projects/nomocrat/doceng26competition/

Marc Tanti
marc.tanti@um.edu.mt
University of Malta
Institute of Linguistics and Language Technology

Stefania Cristina
stefania.cristina@um.edu.mt
University of Malta
Department of Systems & Control Engineering

Alexandra Bonnici
alexandra.bonnici@um.edu.mt
University of Malta
Department of Systems & Control Engineering

Abstract

Develop an OCR model for transcribing images of paragraphs in pages extracted from Maltese PDFs. A train set is not provided and synthetic data must be used. Language resources will be provided to participants to assist with those unfamiliar with Maltese.

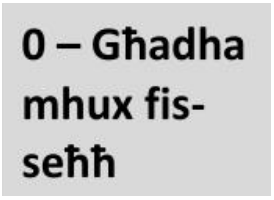
Motivation

A corpus is a large collection of texts that is used for studying language and for the development of large language models, among other things. PDFs are an important source of text for corpora, but they are challenging to extract text from, particularly due to the selectable text in the PDFs being unreliable due to font-based character substitutions (characters appearing as other characters), ligature substitutions, merged columns, and so on. A solution to this is to avoid the selectable text completely and focus on the apparent text by using an OCR (Optical Character Recognition).

Such a system will allow for the extraction of high quality corpora from noisy sources of text, which increases the amount of available data for low resource languages. This creates more inclusion among languages into the digital world. Higher quality corpora also means that less text is needed to get around the noise that would otherwise be present, allowing for language models to be trained with less.

Task description

The Maltese language is the national language of Malta. It is a Semitic language that uses a Latin script for writing. Given that it is a low resource language, this competition is about benchmarking OCR models for Maltese PDFs trained on synthetic data. The OCR is to be applied to rectangular images of single paragraphs only. A development set of gold standard transcribed images of paragraphs from Maltese PDF pages is provided to allow participants to evaluate their models. A test set is held out and only used by the competition organisers to measure the final evaluation metrics of the competition. Below is an example of a data item in the development set:

Image input	Expected output
	0 — Għadha mhux fis-seħħ

The main challenge of this competition is the lack of a train set which needs to be generated synthetically. A second challenge is that the OCR output must be in paragraph form rather than as a list of lines (as some OCRs output). A list of lines is not useful for producing a corpus, and therefore the OCR model must either be trained to output the paragraph as a whole, or the lines must be joined as a post-processing step. Joining lines of text is not trivial as some lines require a space between them and others don't. Some lines end in a hyphenated word that must be made whole again. The Maltese language makes extensive use of dashes (e.g. 'il-kelb' which means 'the dog'), which creates ambiguity for deciding whether the dash must be preserved or treated as a hyphen. A rule-based line joiner for Maltese has been developed and can be found linked below.

Winning criteria

The generated transcriptions are compared with gold standard transcriptions and the CER (Character Error Rate) is measured. The model with the lowest CER is the winner. In the case of a tie, the one with the shortest runtime (on the organisers' computer) is given preference.

Provided resources

- The development set consists of cropped images of paragraphs together with a JSON file giving the ground truth transcriptions in paragraph form. The JSON file will contain both the paragraph form transcription which is what is used for evaluation as well as the list of lines as given in the PDF (including hyphens). Note that some of the paragraphs are in English, which is to be expected in Maltese text as English is very common in Maltese language. There are also some other languages included but very sparsely.

- The evaluation script that can be used on both the test and development sets will be provided (but you will only have access to the development set). The pip requirements file to install will also be provided.
- A list of all characters in both the development and test set will be provided.
- A corpus of Maltese texts can be found here.
- A rule-based Maltese line joiner can be found in the malti Python package.

Evaluation computer specs

The computer that will evaluate the participants' models on the test set has the following specifications:

- O/S: Windows 11
- Processor: Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz
- RAM: 32GB
- GPU: NVIDIA GeForce RTX2080 Ti (11GB VRAM)
- Python: v3.9 with conda v4.12

Rules

- **The deadline for the competition is 30th June 23:59 AoE (Anywhere on Earth).**
- The developed model must be publicly published on HuggingFace. The readme file on the HuggingFace page should provide information about how to use the model as well as how it was trained. Updating the model on HuggingFace after the competition deadline will result in disqualification (newer versions of the model may be uploaded using a new name).
- A Python script containing a class that performs the OCR transcription is to be provided by the participants. The script will be imported in the participants' evaluation code and the class inside will be used. The class must:
    - download the model from HuggingFace and perform any other initialisations in the '`__init__`' function and
    - have a '`transcribe`' function that takes one `PIL.Image` object as a parameter and returns one string containing the transcription of the image. A batch size of 1 is enforced (during evaluation) to minimise the chance of out-of-memory errors.
- The code in the script needs to be clear and readable enough to allow for a manual inspection. Any code that is considered malicious or that appears to gain access to the test set or cheat in any other way will be disqualified.
- Use of the Internet other than to download the model from HuggingFace (such as to access online services) is prohibited.
- The participants can provide their own requirements.txt file that is installed before the one provided by the organiser. The participant requirements.txt file must not conflict or otherwise contradict the package version numbers in the one made by the organisers. Both requirements.txt files will be installed in an Anaconda virtual environment.
- The entire evaluation must install and execute on the organisers computer without any errors in less than 5 hours. The test set is roughly the same size as the development set. All the resources, including the virtual environment, should not require more than 20GB of disk space.