SeanchlÓCR: A platform for Transformer-based transcription of Irish Gaelic Script

Sháhín De Faoite Houshidari*, **Brady Meade***, **Ellen Rushe**
School of Computing, Dublin City University, Dublin, Ireland
{shahin.defaoitehoushidari2, brady.meade2}@mail.dcu.ie, ellen.rushe@dcu.ie

Abstract

This paper presents a tool for interactive digitisation of documents using Irish Gaelic script. The system, SeanchlÓCR, uses a transformer-based OCR model fine-tuned through an iterative refinement process. In addition to achieving lower character error rate than any other currently available system, our proposed framework also provides users with adjusted confidence scores for each word transcribed, allowing users to locate transcription errors and perform post-editing. The framework allows multiple users to post-edit concurrently and integrates semantic search, enabling users to retrieve previously digitised documents. It is hoped that the system can accelerate digitisation of important cultural artifacts by providing an effective transcription, search and post-editing framework[1].

1 Introduction

Seanchló Gaelach (meaning "old Irish script") or *Cló Gaelach*, is a historical typeface used in early printing of the Irish language. Although the typeface has been phased out since the mid twentieth century (Ó Cúlacháin, 1980), there remains a rich legacy of historical Irish documents using this unique script. Unfortunately, due to the scarcity of automated systems to transcribe these documents, compounded by the reliance on specialised, expert transcriptions, many exist in a digitally inaccessible format. This scarcity has lead to a bottleneck in the digitisation of documents of significance to Irish cultural heritage, including Irish language manuscripts, legacy newspapers (such as An Gaodhal, in publication between 1881-1898 (Dereza et al., 2024)), ecclesiastical records, and government documents. Given the historically low-resource status of the Irish language and this typeface, it is hoped that this work will provide insights to other practitioners attempting to build frameworks to generate digitised text using similarly low-resource typefaces.

There are two primary existing works which develop OCR for Cló Gaelach, including the most recent work by Dereza et al. (2024) who developed a BART-based OCR model to transcribe entries from the "an Gaodhal" newspaper, and training data for Tesseract OCR prepared by Scannell et al. (2015). We build upon the insights of these authors to create a novel OCR model which we integrate into a custom interactive framework which provides users with both semantic search and confidence scoring for transcribed characters.

2 OCR for Cló Gaelach

Due to the lack of available dataset matching line-images to captions for Cló Gaelach, we built upon document-images by Scannell et al. (2015), manually splitting these document-level images into line-level images, applying line-level captions to each. The training dataset was prepared in stages. The first dataset consisted of 1,571 images and captions. In order to improve accuracy, we staggered expansion of the dataset, using the most up-to-date version of the model to automatically label documents, before correcting these labels manually. Where possible, we cross-referenced transcriptions against full-document-length transcriptions created by Scannell et al. (2015). The final complete dataset consisted of 2,713 line images and text pairs. In order to split lines and transcribe their contents, we fine-tuned the Tesseract OCR (Kay, 2007) "ga" model on the first dataset containing 1,571 ground-truth line-images and captions. We then fine-tuned the TrOCR (Li et al., 2023) transformer model on the same 1,571 line-images and captions. The system achieves a CER of 3.8% on the dedicated test set. On the same test set, the *An Gaodhal* model (Dereza et al., 2024) achieves a CER of 7.6%

*Both authors contributed equally to this work.
[1]Our video demo in addition to our code/data can be found at this link.

Proceedings of the 19th International Natural Language Generation Conference: System Demonstrations, pages 14–15
October 17-21, 2026 ©2026 Association for Computational Linguistics

while Google Cloud Vision [2] achieves 13.1%.

3 Semantic Search

Due to the low-resource nature of our dataset, we use Irish-to-English machine translation and embed the resulting text using the `all-MiniLM-L6-v2` (Wang et al., 2020) [3]. At document-processing time, the full transcription is translated from Irish to English using Google Cloud Translate primarily, with Helsinki-NLP (Tiedemann, 2020)[4] used as a fallback. The English text is embedded producing a 384-dimensional dense vector. At processing time, the full English translation of each document is embedded and stored as a record. At query time, the translated English query is embedded using the same model, and cosine similarity is used to rank documents by semantic closeness. We note that the semantic search component is presented here as a functional capability. Quantitative assessment would require a gold-standard set of queries with relevance judgements over Cló Gaelach documents, which we leave to future work.

4 Interactive Confidence Scoring

Confidence-guided post-editing is integrated into the framework design. Word-level confidences were calculated by averaging the output probability of the most probable token on a character-by-character basis, for each character in each word. However, this method inconsistently matched incorrect words to low-confidence words, indicating a lack of model calibration. We therefore adjusted this scoring using an offset, calculated using the empirical character-level errors. More precisely, for each transcribed word that does not match its corresponding ground truth, we extract the erroneous characters. We then count the number of times each character has been incorrectly transcribed across the entire evaluation dataset. We denote the number of times a given character c has been incorrectly transcribed as c^E. We then count the number of time c occurs in the dataset, denoted by c^D. The confidence offset is is simply the ratio between c^E and c^D: $offset_c = \frac{c^E}{c^D}$

The output probability is adjusted by simply subtracted the offset for the predicted character. That is, for a given prediction of character c, denoted as $pred_c$ the adjusted confidence $conf_c$ is defined as $conf_c = pred_c - offset_c$. We calculated CER on a character-by-character basis using the evaluation dataset. Characters could then be mapped to their error prevalence. Figure 1 shows the confidence-guided post-editing framework in practice.

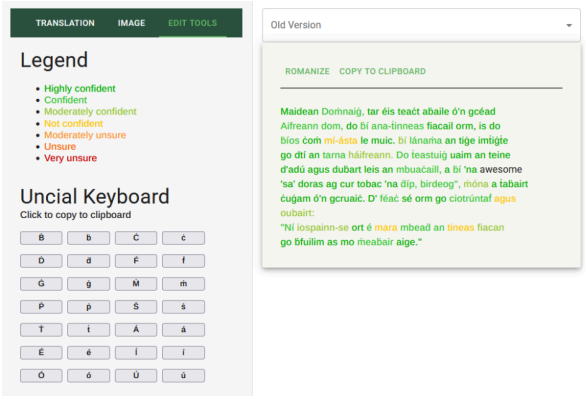


Figure 1: Confidence-guided post editing interface.

[2] https://docs.cloud.google.com/vision/docs/reference/rest

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[4] https://huggingface.co/Helsinki-NLP/opus-mt-ga-en

References

Oksana Dereza, Deirdre Ní Chonghaile, and Nicholas Wolf. 2024. "to have the 'million'readers yet": Building a digitally enhanced edition of the bilingual irish-english newspaper an gaodhal (1881-1898). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 65–78.

Anthony Kay. 2007. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Kevin Scannell, Jim Regan, and Kevin Damazyn. 2015. kscanne/tesseract-gle-uncial: Data and scripts to train tesseract ocr model for old irish fonts (seanchló).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

C. Ó Cúlacháin. 1980. *Tobar na Gaeilge*. Helicon Teoranta, Baile Átha Cliath.