Advancing OCR and Word Sense Disambiguation for the Jawi Script using LLMs and VLMs

Miguel Escobar Varela[a], Faizah Zakaria[b], Seng Guo Quan[c], Ganesh Neelakanta Iyer[d], Pratik Kamarkar[d] and Setphane Bressan[d]

[a] Centre for Computational Social Science and Humanities, NUS

[b] Department of Southeast Asian Studies, Faculty of Arts and Social Sciences, NUS

[c] Department of History, Faculty of Arts and Social Sciences, NUS

[d] Department of Computer Science, School of Computing, NUS

Abstract

In Southeast Asia, many key historical records are written in Jawi, an Arabic-derived writing system historically used for the Malay language. This paper presents a novel approach to making these documents machine-readable through two main contributions: a high-quality OCR system that outperforms previous solutions with a Character Error Rate (CER) of 8.66%, and a context-aware word sense disambiguation model that achieves 99.2% accuracy. We introduce novel datasets and fine-tuned models for both tasks, advancing the accessibility of Jawi documents and enabling downstream NLP applications.

1. Introduction

For over five centuries, the Jawi script - an adaptation of Arabic writing - served as one of the primary means of written communication in Southeast Asia. This project addresses the challenge of making historical Jawi texts machine-readable through OCR and post-OCR correction, contributing to a growing body of Digital Humanities work on historical writing systems (Strange et al, 2014; Hill and Hengchen, 2019). As an under-resourced script, Jawi presents unique challenges and opportunities that we explore through this case study. To contextualize our approach, we begin with a brief history of the script.

Malay, which is an official language in Malaysia and Singapore, and is used elsewhere in Southeast Asia, is now typically written using the roman script, or Rumi Malay. However, for much of the fourteenth to the early twentieth century, Jawi was the script most widely used in written records, letters, manuscripts and tombstone inscriptions. Letters in Jawi largely correspond to those in Arabic, with modifications made to accommodate sounds absent in Arabic such as /ng/ and /p/. Unlike the Arabic script, Jawi largely developed without an institutional center that regulated the form of the script, thus leading to a lack of uniformity in spelling and orthography (Rashid and Juhari, 2006; Collins, 1998; Fogg, 2015). This was especially evident in the use of vowels. A famous example, used by Jawi expert Mulaika Hijjas (2024) is the form بنتڠ)bntng), which can be interpreted in at least five different ways:

banteng (wild ox), *banting* (throw down), *bentang* (spread out), *benteng* (fortification), and *bintang* (star). The absence of diacritics may also result in idiosyncratic use of vowels such that a single Rumi spelling may result in multiple Jawi renderings, for instance, *merah* (red) which can be written مره or ميره.

Vocabulary lists in Rumi began appearing in the early 16th century as European explorers began learning the Malay language (Bausani, 1960; Teeuw, 1959). Outside the colonial administration, however, Jawi was still widely used. Malay printing presses published their books, pamphlets and newspapers in Jawi, retaining its use in the public sphere (Lent, 1978). However, after decolonisation, there was a clear pivot towards Rumi, and Jawi is now rarely used.

Given the historical importance of Jawi, and the current dominance of Rumi, we aim to be able to accurately represent historical Jawi texts in machine-readable formats, in both Jawi and Rumi. This two-pronged approach will enable more widespread access to historical texts in Rumi (as less people read Jawi today), but also preserve the original Jawi script for those who wish to engage with historical documents in their authentic form. We also wish to enable downstream NLP tasks (such as NER, classification and information retrieval) in both scripts.

To achieve these larger goals, we first need to advance the following tasks (represented visually in Figure 1):

1. Accurately render Jawi texts in UTF-8 (through OCR).
2. Transliterate UTF-8 Jawi inti UTF-8 Rumi, using a combination of:
    2.1. A rule-based system (which would render بنتڠ as “bntng”.
    2.2. A disambiguation system, which would convert forms such as “bntng” into the correct words (“bentang”, “bintang”, etc.).

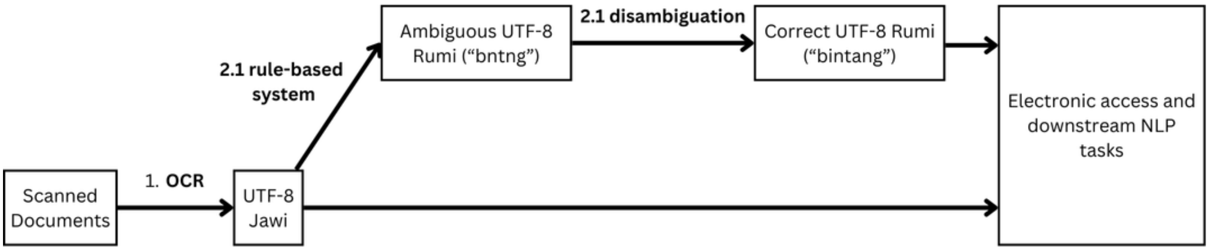
Ambiguous UTF-8 Rumi (“bntng”)
2.1 disambiguation
Correct UTF-8 Rumi (“bintang”)
2.1 rule-based system
Electronic access and downstream NLP tasks
Scanned Documents
1. OCR
UTF-8 Jawi

Figure 1. A diagrammatic representation of our research pipeline

Task 2.1 is relatively straightforward, so in this paper, we focus on task 1 and task 2.2, which present several challenges. While there is a wealth of resources for Arabic OCR, these systems don’t work well for Jawi. The second task, transcribing Jawi into Rumi, also presents unique problems. As noted earlier, the script typically omits short vowel markers, leading to significant ambiguity in interpretation. These characteristics, combined with the historical nature of many Jawi documents, and changes in spelling conventions, make traditional approaches for OCR and transcription inadequate for the Jawi context. For example, Arnia et al (2019) report a 94% accuracy rate in detecting individual Jawi letters using a relatively straightforward algorithm (which they call the Tree Root algorithm). Razali at al. (2024) were also reportedly able to achieve a 94.9% accuracy ret using RestNet34 on individual Jawi letters. However, the problem of these approaches is that they don't scale well to longer texts.

Recent advances in Large Language Models (LLMs) and Visual Language Models (VLMs) have opened new possibilities for OCR and post-OCR correction (Thomas et al, 2024; Madarász et al. 2024; Mattingly, 2024) in under-resourced languages. However, the application of these techniques to Jawi has been limited by the scarcity of high-quality training data, evaluation benchmarks and the unique challenges posed by the script's uniqueness and inherent ambiguities. To address these limitations, we introduce datasets and models for OCR word sense disambiguation.

2. Data and methods

As part of our work, we developed two novel datasets for evaluating OCR and disambiguation tasks in Jawi. The first dataset pairs images of Jawi text with their UTF-8 transcriptions, enabling direct evaluation of OCR accuracy. For this dataset, we used a scanned copy of the *Warta Malaya* newspaper, the first of the independent Malay dailies of the 1930s, and which ran for 10 years (National Library Singapore, n.d.). A selection of scanned pages from this newspaper was provided to us by the National Library of Singapore. We extracted all sentences from these pages and captured each one as a digital image with a corresponding UTF-8 transcription (a combined total of 1139 sentences). It should be noted that these examples sometimes include Roman-script English text, reflecting the multilingual nature of historical Malay newspapers.

Several pages from the *Warta Malaya* had been previously transliterated into Rumi, as part of the Malay Transliteration Project at the National University of Singapore (Emmanuel and Barnard, 2009). Hijjas has also led another project for transliterating Jawi into Rumi (From The Page, n.d.). However, to the extent of our knowledge, there are no openly accessible datasets that include Jawi texts and their UTF-8 representations. Our data is available on Hugging Face at https://huggingface.co/datasets/mevsg/Jawi-OCR-v1.

The second dataset addresses word sense disambiguation, focusing on common ambiguous cases in Jawi script. We collected 2,500 contextual examples centered around the aforementioned Jawi form بنتغ)bntng), and five possible interpretations: banteng (wild ox), banting (throw down), bentang (spread out), benteng (fortification), and bintang (star). Each interpretation is represented by 500 examples, providing a balanced dataset for training and evaluation. These examples were synthetically generated using Open AI's gpt-4o-mini model via their API. The data is available on Hugging Face at https://huggingface.co/datasets/mevsg/bntng-disambiguation-v1.

Both the transcription and the disambiguation datasets were independently verified by two annotators. Our OCR system builds upon the Qwen2-VL-2B-Instruct model, which we fine-tuned for Jawi script recognition using the transcription dataset described above. For finetuning this model, we were inspired by Mattingly's (2024) approach. This model was finetuned on a H100 GPU for 6 hours. The resulting model, and usage instructions are available at https://huggingface.co/mevsg/qwen-for-jawi-v1.

Our word sense disambiguation system builds on recent advances in large language models for Southeast Asian languages. AI Singapore developed the SEA-LION model (2024) from scratch, specifically to work with these languages. For our word sense disambiguation task, we fine-tuned *sea-lion-7b-instruct-research*, an instruct-tuned variant of this model. Our finetuning script ran for 16hours on a single A6000 GPU and our model, together with usage instructions, is available at https://huggingface.co/mevsg/bntng-dis-v2.

5. Results and Discussion

Our OCR system achieved the following error rates:

- Character Error Rate (CER): 8.66
- Word Error Rate (WER): 25.50

This is a significant improvement over Parichuri's (2024) Surya, a state-of-the-art general-purpose OCR model which reports a >95% accuracy for Arabic, but performs poorly on our Jawi data:

- Character Error Rate (CER): 70.89%
- Word Error Rate (WER): 91.73%

Our character and word error rates, though far from perfect are remarkably encouraging given the extremely small dataset we used for training (911 sentences, 80% of our dataset). Given the complexity and multilingualism of these examples, our results demonstrate that VLMs are excellent candidates for future OCR exploration. Our word sense disambiguation model achieved 99.2% accuracy on the validation set. While accuracy is extremely high, we should note that the dataset only included a single ambiguous form, and more research is needed to expand this to a general disambiguation model. That said, this approach is clearly promising given the high accuracy achieved in this initial experiment.

6. Conclusion

Our work convincingly demonstrates the effectiveness of modern deep learning approaches in handling the complexities of Jawi script processing. The combination of improved OCR capabilities with robust word sense disambiguation represents a significant step forward in making historical Jawi texts more accessible to researchers and the public.

References

Arnia, Fitri, Khairun Saddami, and Khairul Munadi. “Moment Invariant-Based Features for Jawi Character Recognition.” *International Journal of Electrical and Computer Engineering (IJECE)* 9, no. 3 (June 1, 2019): 1711–19. https://doi.org/10.11591/ijece.v9i3.pp1711-1719.

Bausani, Alessandro. “The First Italian-Malay Vocabulary by Antonio Pigafetta,” East and West, 11:4 (1960), pp. 229-248.

Collins, James T. *Malay, World Language: A Short History*, (Kuala Lumpur: Dewan Bahasa dan Pustaka, 1998).

Fogg, Kevin. “The standardisation of the Indonesian language and its consequences for Islamic communities,” *Journal of Southeast Asian Studies*, 46:1, (2015), pp. 86-110.

From The Page. “Jawi Transcription Project.” Accessed November 20, 2024. https://www.fromthepage.com/mulaika/jawi-transcription-project.

Gallop, Annabel Teh. “Early Malay Printing: An Introduction to the British Library Collections.” *Journal of the Malaysian Branch of the Royal Asiatic Society* 63, no. 1 (258 (1990): 85–124.

Gallop, Annabel Teh, Wan Ali Wan Mamat, Ali Akbar, Vladimir Braginsky, Ampuan Hj Brahim Bin A.H. Tengah, Ian Caldwell, Henri Chambert-Loir, et al. “A Jawi Sourcebook for the Study of Malay Palaeography And Orthography.” *Indonesia and the Malay World* 43, no. 125 (January 2, 2015): 13–171. https://doi.org/10.1080/13639811.2015.1008253.

Hijjas, Mulaika. “Is Jawi Islamic?” In *Malay-Indonesian Islamic Studies*, 269–93. Brill, 2022. https://brill.com/edcollchap/book/9789004529397/BP000010.xml.

———. “Marsden’s Malay Manuscripts: Reassessing a Colonial Collection.” *Philological Encounters* 8, no. 1 (2022): 38–72.

———. “Teaching Jawi in the Pandemic.” *Teaching the Codex* (blog), February 23, 2022. https://teachingthecodex.com/2022/02/23/teaching-jawi-in-the-pandemic/.

Hill, Mark J, and Simon Hengchen. “Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study.” *Digital Scholarship in the Humanities* 34, no. 4 (2019): 825–43. https://doi.org/10.1093/llc/fqz024.

Lent, J.A. “Malaysia’s National Language Mass Media: History and Present Status,” *Southeast Asian Studies*, 15:4, (1978), 599-612.

Madarász, Gábor, Noémi Ligeti-Nagy, András Holl, and Tamás Váradi. “OCR Cleaning of Scientific Texts with LLMs.” In *International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs*, 49–58. Springer Nature Switzerland Cham, 2024. https://library.oapen.org/bitstream/handle/20.500.12657/93249/1/978-3-031-65794-8.pdf#page=60.

Mattingly, William. “Qwen2-vl-Finetune-Huggingface.” Python, November 26, 2024. https://github.com/wjbmattingly/qwen2-vl-finetune-huggingface.

National Library Singapore. “Warta Malaya.” Accessed November 20, 2024. https://www.nlb.gov.sg/main/article-detail?cmsuuid=a8edfbe8-a5e8-4c08-8b3f-2908fcf56e2f.

Paruchuri, Vik. “Surya.” Python, November 30, 2024. https://github.com/VikParuchuri/surya.

Rashid, Melebek Abdul and Moain Amat Juhari, *Sejarah Bahasa Melayu*, (Kuala Lumpur: Utusan Publications and Distributors, 2006).

Strange, Carolyn, Daniel McNamara, Josh Wodak, and Ian Wood. “Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers.” *Digital Humanities Quarterly* 8, no. 1 (2014).

Teeuw, Andries. “The history of the Malay language: A preliminary survey,” *Bijdragen tot de taal-, land- en volkenkunde,* 115:2, (1959), pp. 138-159

Thomas, Alan, Robert Gaizauskas, and Haiping Lu. “Leveraging LLMs for Post-OCR Correction of Historical Newspapers.” In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, 116–21, 2024. https://aclanthology.org/2024.lt4hala-1.14/.