A Roadmap for the Automated Production of Enriched Reading Environments for Historical Arabic Texts

Joseph C. Hilleary
Tufts University, Department of Computer Science

Introduction

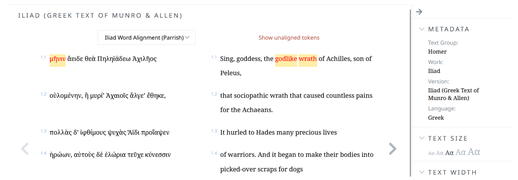
A large portion of the pre-modern documents that have survived to the present day are in Arabic or languages that use Arabic script. These materials are of immense scholarly and cultural importance, but many have been historically inaccessible due in part to the lack of quality open-source digital editions.

One reason for this disconnect has been the lack, until recently, of computational tools for working with Arabic. Despite its importance as one of the most widely spoken languages in the world and its role as both a religious and scholarly language, Arabic has often been neglected in the development of software for natural language processing. In recent years, work by OpenITI, KITAB, and the CAMeL Lab, among others, has begun to change the ecosystem of tools available for Arabic, creating an unprecedented opportunity to make this global heritage more accessible.

This work presents a vision that leverages these new tools to create an automated pipeline beginning with a scanned text and producing the files needed to support the text's inclusion in a modern digital library reading environment.

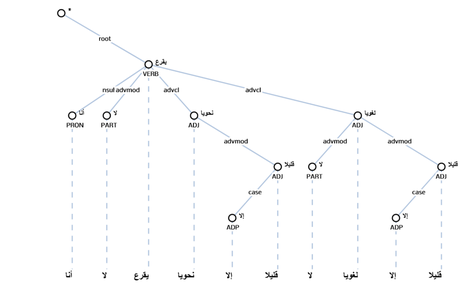
Translation Alignment

Word or phrase-level alignments between a source text and a translation reveal choices made by the translator. They also help to elevate phenomena in the original material that might otherwise be missed by a translation-dependent reader.



Treebanking

Syntax trees, stored in collections called treebanks, show the linguistic dependencies between words in a sentence. Increasingly, the Universal Dependencies system has become a multi-language standard with its claim of language agnostic (or adaptable) dependency relationships.



Modern Ecosystem

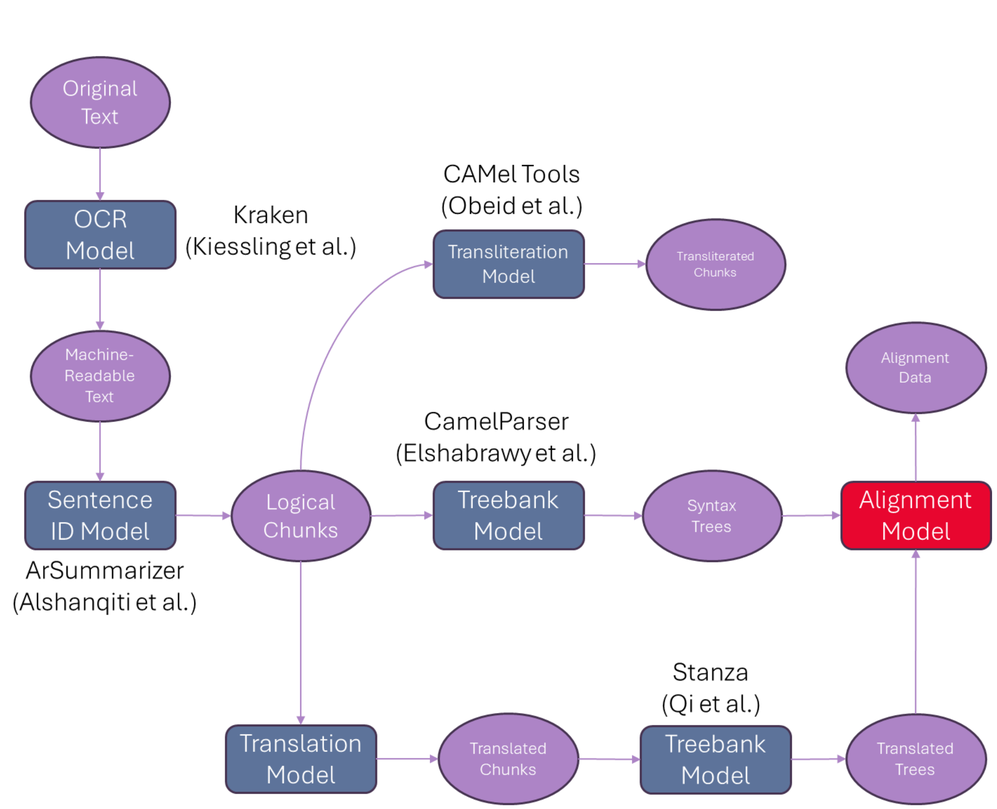
Arabic NLP

- OpenITI and KITAB
  - Digital corpus building
  - Optical character recognition with Kraken (2017)
    - 97% accuracy
- CAMeL Lab
  - Natural language processing with CAMeL Tools (2020)
    - Morphological analysis
      - Disambiguation
    - Transliteration
    - Named Entity Recognition
    - Dialect identification
    - Sentiment Analysis
  - Dependency parsing with CamelParser 2.0 (2024)
    - Support for Universal Dependencies standard
- ArSummarizer (2022)
  - Segmentation of unpunctuated text

Reading Environments

- Perseus Project
  - Beyond Translation (2023)
    - **Translation Alignment**
    - **Treebanks**
    - Token-level annotation
    - Named entities
      - Geospatial annotation
    - Dictionary and grammar support
    - Integration of open-platform annotations
      - New Alexandria
        - Commentaries

Proposed Pipeline



Challenges of NLP for Arabic

- European language bias in tool development
- Linguistic attributes
  - Morphologically rich
    - Concatenative
    - Templatic
  - Orthographic ambiguity
    - Optional diacritics
  - Diglossia and Classical vs. Modern

كتبهم

كُتُبهم or كَتَبَهم

Conclusions and Remaining Work

Most of the tools now exist to build an automated pipeline for developing enriched reading environments for Arabic. While the existing models still have room for improvement, it has become increasingly possible to work with and annotate Arabic texts at scale.

In the short term, there still remains no standard, easily automated tool for aligning translations, current research seeks to fill this gap. In addition, there is potential to build out additional features beyond just dependency trees and translation alignments. In particular, the CAMeL tools NER capabilities could be leveraged if paired with other open-source knowledge bases, such as Wikipedia, to provide information about persons and places referenced in the text.

There also trade-offs to automation. Hitherto such editions have been produced painstakingly by experts and as a result have a high-level of accuracy and fidelity relative to the source material. Linking so many models together also increases their collective potential for error. Nevertheless, this approach provides a significant, if imperfect step toward broad accessibility of a large number of texts.

Acknowledgements

Abdullah Alshanqiti et al. "Employing a Multilingual Transformer Model for Segmenting Unpunctuated Arabic Text" *Applied Sciences*, 2022
Ahmed Elshabrawy et al. "CamelParser2.0: A State-of-the-Art Dependency Parser for Arabic" 2024
Gregory Crane, "Beyond Translation: a reading environment for the next generation Perseus Digital Library." *The Perseus Journal of Data Preservation and Sustainability,* 2023
Benjamin Kiessling "Important new developments in Arabographic optical character recognition (OCR)." *Al-ʿUṣūr al-Wusṭā*, 2017
Joakim Nivre, "Universal Dependencies." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017
Ossama Obeid et al. "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing." *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020
Peng Qi "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." 2020