Next generation AI-based ecosystem for Noolaham Foundation

Project Title	Large Scale Annotation, Storage and Analysis of Digitised Sri Lankan Tamil Content
Project Location	Sri Lanka
Implementing Institution	Noolaham Foundation
Total Budget	
Duration	-
Team	Shaseevan Ganeshananthan, Prashanth Srinivasan, Saatviga Sudhahar

Project Summary

The project aims to do large scale annotation, storage and analysis of Sri Lankan Tamil content. This is related to the field of semantic culturomics in which researchers data mine large digital archives to investigate cultural phenomena reflected in language and word usage. It is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts. The underlying data is from Noolaham Foundation, a Digital Archive and a Digital Library undertaking the critical work of documenting, digitally preserving and providing free and open access to knowledge bases and cultural heritage of Sri Lankan Tamil speaking communities. The archive contains digitized text from Sri Lankan newspapers, books, magazines, pamphlets etc from various sources totalling up to approximately 100,000+ documents. It also includes a web archive and born-digital data in text format which would be included in our pipeline.

Goals

The goal of the project is to build an ecosystem that analyzes large scale Sri Lankan Tamil content using natural language processing methodologies and generative AI.

Objectives

- Converting digitized content from Noolaham project to text with meta-data tagging. This includes converting newspapers, books, magazines, pamphlets from pdfs and images to text. We do Layout analysis of newspapers, Optical character recognition, Building document type storage, XML conversion of text,
- Auto-labeling of documents starting from manual annotation and training a model to automatically annotate content
- User Interface to display extracted text with annotations for all categories in Noolaham's content
- Building language processing resources using natural language processing resources and building Knowledge engineering capabilities.
- Building a generative AI based tool to query Noolaham's content, similar to ChatGPT.

Progress & Achievement

1) Language Pre-processing

- **Layout Analysis**

  All documents available in the digital archive need to be converted to text format via OCR. For

newspaper articles it's important to perform document layout analysis before the image is being sent to OCR. It is the process of identifying and categorizing the regions of interest in the scanned image of a text document. A reading system requires the segmentation of text zones from non-textual ones and the arrangement in their correct reading order.

Progress

Completed development of the digital content conversion pipeline that performs layout analysis for Noolaham's content.

To be done

The pipeline has to be run for all the digitized documents in Noolaham.org.

● **Optical Character Recognition**

All document images need to be converted to text via an OCR system. The text would be displayed in Islandora UI along with all other available file formats.

Progress

Completed development of the digital content conversion pipeline that performs OCR. We used tesseract-ocr (https://github.com/tesseract-ocr/tesseract) as part of the pipeline which is currently being maintained by Google and offers the most accurate results in current benchmarks. For newspapers once the Layout analysis is complete, the resulting chunks of images should be fed to the OCR tool.

To be done

The pipeline has to be run for all the digitized documents in Noolaham.org.

- **XML Conversion**

The OCRed content will be saved and made available in XMLschema based metadata standard formats such as METS or ALTO for future use. The content from the web archive and the born-digital data would be also converted to XML.

Progress

Completed development of the digital content conversion pipeline that does XML conversion in METS format. For example when a document is converted to text, an xml file is created that saves metadata including Title, Author, Place of Publication, Publisher and Date.

To be done

The pipeline has to be run for all the digitized documents in Noolaham.org.

● **User Interface development**

We intended to use Islandora, a collaborative open source framework to manage digitized assets and develop it to showcase digitized content from Noolaham in image, pdf and text formats. The UI will also show metadata and keywords for each document from manual annotations done by annotators.

Progress

In Progress

● **Metadata Creation**

We require certain metadata to be attached to the documents displayed in the Islandora UI. eg: For a newspaper article the meta data would include, the name of the news outlet, date, title and author of the article. This project requires a few human annotators to manually add metadata to all the documents available from Noolaham in the Islandora platform. Additional annotations such as keywords or topics for the documents would also be manually annotated initially.

Progress

Not started yet

- **Auto-labeling**

  We intend to build a machine learning model to auto-label documents with specific annotations such as keywords or topics. We would train a model using annotations performed manually and use the system to auto-label new documents. This would be cross-checked and corrected using a human in the loop which then adds to the training set of the model.

Progress

Not started yet. Can be done when manual annotations are available.

2) Processing Resource Layer

Processing resources refer to resources whose character is principally programmatic or algorithmic, such as tokenizers, chunkers or parsers used to process the Tamil language. This component in the platform would include all necessary tools to parse Sri lankan tamil text such as sentence splitter, tokenizer, POS tagger, dependency parser, Morphological analyser/generator, named entity recogniser, coreference
resolver, pronominal resolver and word sense disambiguation. Figure 2 below shows how a corpus could be processed by a set of processing resources and used for Analytics.

Progress

An initial analysis has been done on existing tools but an uptodate gap analysis has to be done in order to initiate this.

3) Knowledge Engineering

Knowledge engineering would be the ultimate point we would reach when all the previous resources are in place. Using the language and processing resources there is the possibility of analyzing the huge amount of Tamil text data for new insights.

Progress

We are building a Tamil GPT using generative AI techniques that possess all knowledge from Noolaham data and will be able to answer questions regarding that. Currently the core technology is being built.

Conclusion

As part of the language preprocessing layer of the ecosystem we have completed the development of the digital content conversion pipeline. As next steps layout analysis, ocr and xml conversion has to be run on the whole of Noolaham's content to convert it to raw text. This requires significant resources and budget related to server cost and running time. I will submit a separate proposal specifying those requirements.

Separate efforts have to be initiated to build the processing resources layer and the knowledge engineering layer.