The EDEAL Project for Automated Processing of African Languages

Anna Borovikov
CACI
4831 Walden Lane
Lanham, Maryland 20706, USA
aborovikov@caci.com

Eugene Borovikov
CACI
4831 Walden Lane
Lanham, Maryland 20706, USA
yborovikov@caci.com

Bradley Colquitt
CACI
4831 Walden Lane
Lanham, Maryland 20706, USA
bcolquitt@caci.com

Kristen Summers
CACI
4831 Walden Lane
Lanham, Maryland 20706, USA
ksummers@caci.com

Abstract

The EDEAL project seeks to identify, collect, evaluate, and enhance resources relevant to processing collected material in African languages. Its priority languages are Swahili, Hausa, Oromo, and Yoruba. Resources of interest include software for OCR, Machine Translation (MT), and Named Entity Extraction (NEE), as well as data resources for developing and evaluating tools for these languages, and approaches—whether automated or manual—for developing capabilities for languages that lack significant data resources and reference material. We have surveyed the available resources, and the project is now in its first execution phase, focused on providing end-to-end capabilities and solid data coverage for a single language; we have chosen Swahili since it has the best existing coverage to build on. The results of the work will be freely available to the U.S. Government community.

1 Introduction

The Government acquires massive quantities of textual data in foreign languages in the course of various operations. This content must be processed in a variety of ways: identifying important elements of information, translating all or part of the content, making the data available to the appropriate interested parties, etc. Machine Translation (MT) offers a significant contribution to this process. Tool suites to enable processing and analysis of foreign language material can, and in some cases do, use MT and related technologies to enhance and streamline this processing.

Tools for MT and the strongly related areas of OCR and Named Entity Extraction (NEE) are readily available for a variety of languages, although the state of the art in accuracy and maturity of the tools varies according to the language. The suites for exploitation, processing, and analysis of foreign language material have not, however, historically focused on African languages. The EDEAL project seeks to enable this processing for African languages, by identifying suitable existing tools and augmenting where needed and appropriate, as well as collecting data sets suitable for use in developing tools, training statistical tools, and evaluating them. Because many African languages are likely to have comparatively sparse available resources, both in the form of available data and in the form of reference material or ready access to native speakers, the project is also looking for methods of developing capabilities for new languages in the absence of full resources, leveraging the scant resources that may be available in the best way possible.

This project began with a phase of *initial assessment* of the available tools and data suitable for use in the context of the target processing systems, for four priority languages. This was followed by a plan for execution based on the findings, and the project is currently in the middle of its first execution phase, which focuses on the generation of a full corpus and end-to-end processing capabilities for Swahili; additional phases will focus on additional languages, along with methods for leveraging scant resources.

2 Initial Assessment

The initial assessment focused on four identified priority languages: Swahili, Hausa (Latin script and Arabic script), Yoruba, and Oromo (Latin script only). Its goal was to quickly survey the available tools and data for OCR, MT, and NEE for these languages that might be suitable for use in the target document processing systems. The target systems imposed the constraint that the processing must be capable of running as an integrated part of a larger system without human intervention, preferably through an API, although command line calls are also acceptable. As these systems run on Windows, this context also introduced a strong preference for components that run on Windows as well, to minimize the footprint and complexity of deployment.

This initial assessment, completed in January of 2009, identified the following.

Data was the most needed element of the task. We sought highly varied data sets, to represent the types of material that are opportunistically collected, with available ground truth or answer keys for at least one of the types of processing of interest. The particular variety of interest is specified more fully in Section 3.1. We found no such collections of substantial size that were available for projects other than academic research. We did, however, find sources of data that can serve as components of such a set, such as nearly parallel corpora of Swahili proverbs, health information brochures, etc.

OCR was found to be effective with commercial tools for Swahili, Hausa in Latin script, Hausa in Arabic script, and Oromo in Latin script. Evaluation was performed automatically for documents collected electronically, since ground truth was available; for paper documents, evaluation was performed manually by native speakers of the language, who rated the quality on a scale of 1 to 5, with textual descriptions of the meaning of each value. In the case of Oromo, the software does not claim direct coverage of this language, but running standard Latin script OCR worked very well. This same approach did not work well with Yoruba, as the diacritics were not handled appropriately.

MT was found to be effective with available Government tools for Swahili and Hausa. The evaluations were performed by native speakers, rating results on a scale of 1 to 5, with textual descriptions of the meaning of each rating. We did not perform automated evaluation with a metric such as BLEU, and we did not evaluate fully trainable tools by training them for the languages of interest, due to the lack of an appropriate data set for our use.

NEE was not handled for the target languages by any fully available production tool that we found at the time of the initial assessment completion, although multiple projects were in process to at least potentially lead to such a capability. In addition, seed rules for entity extraction for Swahili, Hausa, and Yoruba are available through the REFLEX project; there is no executable interpreter for these rules, but the rules for Swahili and Yoruba are written in a computationally-oriented manner suitable for such an interpreter.

Developing capabilities with scant resources is a topic of a small number of academic projects.

As a result of this assessment, in combination with a Government priority to focus first on end-to-end support for a single language, the first execution phase works with Swahili. This phase includes the creation of a varied corpus of documents, with OCR ground truth, named entity tagging, and sentence-aligned translations for all documents, as well as following up on the initial assessment results to provide deeper evaluation and support for OCR, MT, and NEE in Swahili.

3 Corpus Creation

The EDEAL project is currently engaged in creating a corpus of varied documents in Swahili, with appropriate data for developing and evaluating OCR, MT, and NEE capabilities. The target size of the corpus is 500 pages and at least 100,000 words.

All data collection, annotation, and creation is performed by native speakers of the language, and all translation and annotation is verified by a second speaker.

3.1 Data Variety

A critical element of this corpus is the variety of the data. Since opportunistically collected documents may come from any source, any data set used for developing, training, or evaluating tools for use in this context must represent a variety of types of content and style. Specifically, we are committed to ensuring that the corpus content includes a variety of registers, subject matter domains, and genres, and that the files include a variety of types and sources of noise (typographical errors, lower quality scanned images, etc.). For each document, the following information is recorded:

- Source
- Text generation, whether machine print or hand print
- Script and encoding
- Genre, as defined by Mikhail Bakhtin and presented on Wikipedia.
- Formality of register, as defined by Quirk, et al. and presented on Wikipedia
- Comments about any other noteworthy elements, including quality and errors

We are selecting documents to provide a significant variety in all of these areas. We are also providing a moderate variety of layout formats; however, we are focusing primarily on simple document layouts, sometimes by extracting the main zone on a complex page, in order to increase productivity of the manual creation of the variety of artifacts required for this collection.

3.2 Corpus Artifacts

For each document in the corpus, we are producing the following set of artifacts:

- Original document, in its original format.
- Plain text rendition of the document content. In some cases, this may be identical to the original document.
- Page images. In some cases, this may be identical to the original document, if the source is a paper document or a scanned image of paper.
- Translation, with sentence alignment.
- Entity tagged file, produced by manual tagging with the Callisto tool from MITRE.

The plain text and the page images together form the basis for work on OCR. The sentence-aligned translations provide the basis for work on MT. The tagged files provide the basis for work on NEE; these files are fully tagged, in that they are tagged for all entity mentions, including nominal and pronominal mentions in addition to named mentions, and they are tagged for within-document co-reference relationships among the mentions. Originals are retained for traceability.

At this stage, the project is creating a single version of each artifact. A potential future enhancement is to provide multiple renditions of artifacts for which this adds value; these categories would include at least page images with varying quality, which we have created on a small scale, and multiple reference translations to be used in automated scoring of MT output.

4 Tool Capabilities

The current phase follows up on the initial assessment results for the tools, in multiple ways.

As the corpus becomes available, we are able to perform fuller evaluations of tool accuracy, with this varied set of data, and to identify whether the tools perform differently on documents with different characteristics, among those we are recording, as specified in Section 3.1.

We experimented with augmenting the OCR for Swahili with a dictionary. For clean documents, this made no difference, but with documents where we introduced noise by damaging the paper and then scanning it, the presence of a dictionary increased accuracy from an average of 88% to an average of 90%.

We have produced an interpreter for the computationally represented seed rules for entity extraction available for Swahili and Yoruba. We have also begun exploring trainable approaches to NEE, to produce this capability for the languages of interest, using Conditional Random Fields (CRFs) or the LingPipe infrastructure. In addition, we are tracking the emergence of production-ready NEE capabilities, either with full training capacity or developed specifically for African languages.

5 Future Work

The current execution phase of the EDEAL project will continue to build the Swahili corpus and work with Swahili processing tools as described in Sections 3 and 4. Planned future phases will focus on additional languages, drawing from the initial priority set: Hausa, Yoruba and Oromo, as well as continuing to enhance capabilities from languages in earlier phases; as new tools and methods emerge, the project will consider them. As a specific example, any newly emerging tools for Swahili entity extraction that do not become accessible to the project in time for inclusion in this first phase will be considered in the next phase, although that phase is expected to focus primarily on the Hausa language.

In addition, future phases will pursue approaches to developing capabilities in the presence of scant resources, with the goal of leveraging and expanding current available academic work into a comprehensive and production ready approach that can make the best use of whichever types of resources are available for a particular language and incrementally add the value of additional resources as they become available.

Acknowledgments

The work described here is performed by a team at CACI that includes, in addition to the authors, Marta Cruz, Mark Turner, and a large team of native speakers of different African languages.

This work is sponsored by funding from the Defense Intelligence Agency (DIA) under contract GS-35F-0342N. We are very grateful for the wise guidance of Nick Bemish and Theresa Williams.