Language Processing for Analysis and Investigation

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

Diane Chandler
CACI
4831 Walden Lane
Lanham, MD 20706
dchandler@caci.com

Abstract

This paper describes an operational case and document management and exploitation system, GlobalView, that includes Machine Translation (MT) for use as an aid to human effort in analysis and investigation. It also presents the REFLEX platform for experimenting with language processing tools.

1 Introduction

We present the use of Machine Translation (MT) software within the GlobalView document and case management suite and within the REFLEX (Research on English and Foreign Language Exploitation) platform. GlobalView manages documents collected in connection with a case, along with derived artifacts, and it provides a workflow that incorporates automated processing tools, including MT. This system is used in investigative operations. The incorporation of MT has formed a critical element of the system's success, and the results of MT have made possible concrete instances of mission success.

2 GlobalView

The GlobalView software system ingests documents of all types, organized into cases, and provides a full-featured workflow that includes a variety of third-party automated language processing tools, in addition to offering core document management support such as version control, metadata entry, and search. The automated processing includes Optical Character Recognition (OCR) for deriving text from scanned image documents, extraction of text from electronic documents, MT from a wide variety of languages into English, and Named Entity Extraction (NEE) and transliteration, as well as processing required to connect these elements, such as transforming character encodings as required. Both MT and NEE and transliteration can contribute to making foreign language material comprehensible to non-linguists, MT by providing an indication of the subject matter and NEE by providing comprehensible representations of the people, organizations, and other such entities referenced in the content.

The artifacts produced by each processing step are stored together with the original file, in a single virtual folder with appropriate metadata attached. Thus, when an analyst accesses a document, that user automatically has direct access to all outcomes of applied automated processing, including MT and NEE results when available. All manually created artifacts, such as manual translation, are stored in the same way.

Additionally, all English content, including translations, is indexed for searching. For example, a search for an English term such as "nuclear weapon" might return results of a variety of types: documents in English, where the term was found in the original; documents in various foreign languages, where the term was found in human-entered metadata or in a human translation; and documents in various foreign languages that may not have received attention from a human linguist yet, where the term was found in MT output. MT thus both contributes to the ability to find a docu-

ment via search and contributes to the available information about documents identified in other manners.

GlobalView is designed to support the use of MT in determining priorities for human translation and in indicating the relevance of a document to an investigation, by offering the results as a rough indication of document subject matter, without competing in any way with the process of human translation to provide a full representation of the document content. MT can also be used in GlobalView as a starting point for full human translation when applicable, and this approach has been used in operation for a set of Spanish documents. Although the translating linguist needed to perform a significant amount of correction, the MT output did provide a head start on processing the document set.

For example, consider an investigation that produces thousands of pages worth of content in a language for which a large number of linguists are not available. The automated workflow in GlobalView can process these pages, producing the kinds of content indication described above, without requiring human intervention except to scan the paper and provide basic metadata for the files. At this point, linguists can browse and search the case, identifying the files that appear to discuss topics of high relevance to the case, either based on the occurrence of keywords or based on conceptual, or "theme" indexing that identifies topics indicated by those keywords, and also identifying files that mention entities of particular interest. They can focus their initial translation efforts on these documents, or the parts of these documents where the indicators of relevant content occur, allowing analysis of the most fruitful case elements to proceed while the rest of the documents are still undergoing full translation. They may also choose to use the MT output as a starting point for full translations, for the documents where this will speed their work. These uses of automated processing to augment human effort streamline the case processing workflow and increase the capabilities for successful completion.

GlobalView has provided the basis for concrete successes related to the customer's mission. For example, GlobalView enabled the full exploitation of over 10,000 pages of foreign language investigative data over a three-day period. GlobalView has been used to cross-reference counterterrorism data with information captured in foreign hot spots. It enabled linguists to link two insurgents who were already jailed in a foreign country with a money laundering case in the United States.

GlobalView also faces some challenges that are particular to an investigative environment such as this customer's. For example, many documents, including the set of more than 10,000 described above, are scanned from paper, meaning that the system performs OCR prior to MT, providing MT with input that contains any errors from OCR. In many cases, due to the quality of the original paper, this OCR output includes significant errors, providing MT with highly noisy input data. Additionally, many of the items treated as individual documents for translation in this system are extremely long, consisting of many distinct logical documents combined together; this tests the scalability of all the automated tools and can create incorrect apparent contexts for content at the boundaries of logical documents. Although MT can prove useful in aiding document exploitation, these data characteristics limit its applicability, and it is not valuable, nor is it applied, to all data in the system.

3 REFLEX Web Platform

As the field experiences described above show, the expected effectiveness of MT will vary significantly with the data and the use for which the output is intended. Under the REFLEX program, CACI has developed a Web platform that provides a means for government users to experiment with MT and other language processing tools. Users can access the system through a browser interface, upload their own data sets, run experiments with the available tools that they select, and store their results.

Figure 1 shows a sample screen shot of the REFLEX interface; this particular part of the interface is used for creating a new test that runs a batch through a selected series of steps.
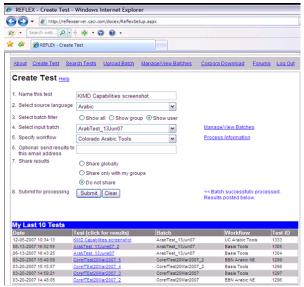
Figure 1. REFLEX platform screen shot.

The language processing tools in the REFLEX web platform include commercial MT engines and research translation software, as well as other language processing components, including NEE. Among its sources of value, this platform offers a means for government users to become familiar with the state of the art in commercial products and offerings that may be coming soon from current research, in order to help inform decisions about whether and how to incorporate language tools into their own processes. REFLEX can therefore assist users in the process of planning to achieve the maximum effectiveness from these tools.

Acknowledgments