

Machine Translation for Triage and Exploitation of Massive Text Data

James E. Andrews

National Ground Intelligence Center
2055 Boulders Road
Charlottesville, VA 22911

james.e.andrews1@us.army.mil

Kristen Summers

CACI
4831 Walden Lane
Lanham, MD 20706

ksummers@caci.com

Abstract

The National Ground Intelligence Center (NGIC) collects massive quantities of textual data in foreign languages. To support exploitation in light of intelligence requirements, a triage process must be applied to this data as those requirements emerge, to identify the most useful data for further exploitation. Machine translation provides critical support for this triage. This paper outlines the types of collected data and the different challenges they present for machine translation, as well as the types of triage to support for collections of this nature, and the issues raised for machine translation by those uses.

1 Introduction

The Government acquires massive quantities of textual data in foreign languages in the course of intelligence operations. The Harmony program at the National Ground Intelligence Center (NGIC) collects such data. It collects data through a suite of deployable tools in the field, including the full-featured Deployable Harmony DOCEX Suite (DHDS), with an Oracle database and full workflow; the comparatively lightweight DHDS-CT, available with lighter databases and without workflow; and the extremely lightweight DHDS-T for processing individual documents. NGIC maintains the national Harmony database serves as a central repository for data collected through these deployable tools and other means. The quantity of available data far exceeds the capacity of human linguists to fully translate, or even review for relevance to current information needs or likely in-

telligence value in the future. Therefore, automated tools must play a critical role in managing this data and empowering personnel at all steps in the life cycle, from warfighters to long-term intelligence analysts, to effectively exploit the information derivable from that data that is relevant to their current needs. Machine translation forms an essential element of this processing, whether it is used directly, to provide an indication of the general meaning of a piece of text for individuals who do not speak the original language, or indirectly, as part of a suite of tools to perform an automated contribution to triage.

The overwhelming quantity of documents and the need for automated tools to enable their effective exploitation will increase significantly with the fielding of the “Dirty-to-Clean” (D2C) tool that produces clean representations of documents from captured electronic media, enabling their import into and dissemination across multiple secure national networks from captured hard drives, thumb drives, CDs, DVDs, etc. A typical hard drive may easily contain 60 GB of data, ranging from conventional office documents through multimedia files such as audio messages and video files, to files such as video games, cached web pages, etc. Even without the increased breadth of data types, the simple increase in quantity creates a need for automated processing. At an average of 50 KB per bitonal scanned page (Wikipedia, “Tagged Image File Format”), a collection of over 1.2 million document pages is required in order to reach the quantity of data in one such hard drive. Thus, simple electronic collections quickly outstrip the scale of even massive, large-scale paper scanning efforts. Kryder’s law dictates that this divide will only increase, making the need for automated processing of the collected data ever more acute.

This paper defines types of acquired foreign language written text data that raise different considerations and needs for machine translation and associated automatic processing, and it also defines a set of goals for automatic processing in the contexts encountered by the NGIC Harmony program. It discusses issues encountered with Machine Translations with respect to these data types and needs, and priorities in that context.

2 Acquired Foreign Language Data Types

Collected documents that consist of text in foreign languages fall into three main categories:

- Paper documents are acquired in hard copy and then scanned as part of their collection process. In this case, the content of the text is not directly available without interpretive processing, such as OCR. The quality of the original paper ranges widely, but the scanning process is under the control of the Government organization performing this process, so settings such as automatic image cleanup, dots per inch (dpi), etc., can be standardized.

Voluminous examples of such data are found in large-scale scanning efforts throughout OIF/OEF related activities. The goal of these efforts was to scan as much data as possible, as quickly as possible. To this end, much of the scanning was performed at a low resolution and with minimal attention to matters that affect later processing, such as correct orientation of pages, the condition of the scanner plate, etc. Even with these challenges, an experiment on a sample of 2,000 files found 137 documents that successfully completed OCT and MT processing, and documents were identified within this set that matched queries intended to reflect current Priority Intelligence Requirements (PIRs). It is reasonable to predict that with consistent application of best scanning practices and scanning at a minimum of 300 dpi, a high percentage of the documents would be successfully processed automatically.

- Captured electronic image documents are acquired in electronic form, but the electron-

ic form represents an image, such as a scanned paper document or a screen shot, rather than directly encoding the text itself. As with paper documents, the content of the text is not directly available without interpretive processing, such as OCR. Unlike paper documents, the process of capturing images of these documents, of course, remains entirely uncontrolled; the images may have been produced in absolutely any manner.

For example, the use of digital photography has expanded dramatically over the last decade. Even most cell phones today are equipped with built-in cameras with varying levels of pixilation. Images of signs, documents, notes and other writings are being found with higher frequency on memory sticks, thumb drives, CDs, laptop hard drives. The files are often JPEG images or even small MPEG videos, typically with less than high-grade resolution.

- Captured electronic text documents are acquired in electronic form and directly represent their textual content. Data of this type includes Word documents, Excel documents, e-mail, etc. In this case, the text to translate can be made directly available to machine translation software, but even under these circumstances, details of formatting and extraction of the text can pose challenges, in addition to those presented by the content itself.

For example, in the case of MS Word or Excel spreadsheets, an extensive use of tables and bullet formatting, wrap text and pop-up comments, poses great difficulties for text extraction tools, such as those used in the D2C. The tools are not able to properly render or allow the user to reconstruct the syntax of the original document contents, thus degrading the machine translation output accuracy and not reflecting the true content of the document or spreadsheet.

3 Goals of Automatic Processing

For the purposes of the Harmony user community, the goals for machine translation and related auto-

mated processing typically focus on *triage*, distinguishing documents that contain information with value, or potential value, from those with no relevance to current needs. The personnel performing initial document review frequently possess limited to no skills in the original language of the documents; likewise, analysts may not read the original languages of the documents they must consider. Therefore, automated aid is required for initial reviewers to identify significant documents; such identified high-priority can then be passed to senior linguists for deeper review, “gisting” (creation of quick summary translations), translation, and reporting, to form the basis for the next steps of exploitation and analysis.

Since intelligence needs change continuously, in response to changing situations and new information and interpretations of that information, the triage determination must be made repeatedly, with new criteria and new concerns each time. Therefore, no single triage process, tuned for a set of particular requirements, can suffice; general mechanisms must be applied repeatedly with different particular requirements.

Machine translation supports these needs by enabling three facets of triage determinations, each of which may be used alone or, preferably, in combination with the others.

- *Content-based* triage decisions can be made by searching for keywords, or other retrieval-like activities. (Whether these activities are executed as search on a collection or filters as documents flow through a system does not alter the basic nature of the task). For these purposes, the task is essentially a specialized variant of Cross-Language Information Retrieval (CLIR), and machine translation plays the same role that it does in CLIR: in principle, it can either translate an English representation of a query to run on the original content or it can translate original documents in order to support English queries. The latter is the more common case and the one supported by the Harmony deployable tools.
- *Metadata-based* triage decisions make use of metadata about the files or their content in order to identify documents of interest. The metadata can reflect information that is available at the batch level, such as the loca-

tion of collection, or information that requires consideration of each document individually, such as a listing of the names of organizations referenced in the document. Machine translation can play a role in increasing the quantity of the latter type of data that is available, without requiring human examination of each individual document. For example, lists of automatically identified organization names will prove far more valuable to English-speaking reviewers if those organization names are transliterated according to a community standard (which enables repeatable retrieval) and/or translated to canonical English forms when such forms are known.

- *Human consideration* triage decisions are made when a human briefly reviews a file in order to determine its significance or relevance to current intelligence needs. This element of triage typically occurs after candidate documents are identified by retrieval or other selection on the basis of automatically available data, using one or both of the above methods. Machine translation of the original files can often produce sufficiently descriptive and meaningful English content to support this activity, even when its results are far too rough to serve as a direct English representation of the original.

Automatic processing in Harmony deployable tools therefore seeks to serve the needs outlined by the elements above. Translating the documents supports both content-based triage and human consideration by non-linguists. This and other processing can contribute to available values for metadata-based triage, depending on the particulars of the extraction of individual metadata field values.

4 Paper Documents

With documents acquired as paper, the major challenge for machine translation, as well as other automated processing, lies in the lack of direct representation of the textual content. For machine print documents, this content can be interpreted by an OCR engine. Although prior work has shown small quantities of randomly distributed OCR errors to have minor effects on direct retrieval in the

original language (Taghva, et al., 1994), the accuracy of OCR has a much more crucial role in effective automatic translation and triage of captured documents. Machine translation is typically highly sensitive to the details of its input and does not naturally make the use of repetition and partial matches that reduces the impact of OCR errors on retrieval. Even the process of producing aligned data for use in machine translation requires special handling in the presence of OCR errors (Dagan, et al., 1993); similarly, our own work with the Army Research Lab (ARL) on the effects of OCR errors on Named Entity Extraction, which uses similar linguistic information to MT, although generally requiring less sophisticated detail, suggests that the text processing accuracy drops off roughly linearly with the OCR accuracy. The Government has recognized the sensitivity of Machine Translation to previous processing such as OCR, as evidenced by the task definition in the Multilingual Automatic Document Classification Analysis and Translation (MADCAT) program under the Defense Advanced Research Projects Agency (DARPA), which measures the accuracy of printed or handwritten text recognition by the accuracy of machine translation of the resulting text.

In addition, the level of OCR errors encountered in collected document images often far exceeds the range usually considered for purposes of mass-market document retrieval. This occurs both because OCR remains at a less mature stage for some languages of interest than for commonly studied retrieval languages such as English and others common in Western Europe and because the paper documents are often collected in poor condition, limiting the accuracy and detail of the available information for OCR software to consider. Additional challenges are encountered when the DOMEX activity scans the images at a very low DPI or in gray scale when color would preserve greater relevant information, in order to reduce file size because of limited storage or to facilitate quicker data transmission. The end result is an image that is not conducive to current machine translation tools, even after the laborious application of current image clean-up and magnification tools.

The effectiveness of OCR depends significantly on the quality of its input images. Therefore, a high-quality document preparation and scanning process is necessary in order to produce the most

meaningful and usable results for later stages. Although, as noted above, the condition of the original paper lies beyond the control of document exploitation teams, the potential to represent that original as usefully as possible in an image for use by OCR is generally available.

The extremely large collection described above offers an excellent example of this phenomenon. The condition of the original paper documents formed a factor beyond the control of the collection and scanning process, but careful scanning at a high resolution can produce images that are both more directly useful and more amenable to improvement with automated tools than scanning at a low resolution and without attention to matters like the condition of the scanner plate. As an illustration of these differences, Figure 1 and Figure 2 show a portion of a sample document scanned similarly to this collection and a 200% zoom of a fragment of that portion, respectively; Figure 3 shows a 200% zoom of a similarly sized fragment of a document scanned carefully at 300 dpi. (Note that the genuine original scans are lower quality than the rough example in Figure 1.) Although Machine Translation was able to process most of the files that completed OCR without errors and with some text output, the OCR challenge and low quality of the consequent input remained the main challenge for Machine Translation.

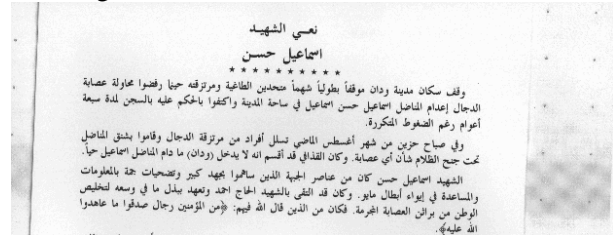


Figure 1: Sample document fragment.

مرزقة الدجال وقاصوا

Figure 2: Small portion of document in Figure 1, at 200% zoom.

أضع برنامجاً أسلي

Figure 3: Small portion of document scanned at 300 dpi, at 200% zoom, for comparison.

Note that for handwritten material, solutions for recognition of the full content do not approach the quality of OCR software for machine print. Therefore, with such input, it is less likely that machine translation will have significant incoming data to process; partial solutions, such as word spotting,

are more naturally suited to use with static glosses than with full machine translation. Again, current Government efforts reflect a recognition of this issue; for instance, the MADCAT program has a focus on handwriting recognition to provide a basis for machine translation, addressing the largest current impediment to automatic exploitation of such documents.

Even with extremely large collections, the quality of images and OCR output outweighs volume as the primary source of the challenge for paper documents. Since each document must be scanned into image format with human effort, the total quantity of documents for later processing is naturally limited. In addition, although the magnitude of existing paper collections is enormous, such collections are typically dwarfed by electronic repositories, if only because of the ease of replication within the latter medium.

5 Electronic Image Documents

With document images acquired electronically, the challenge of deriving the content via OCR remains as for paper, but its difficulty is increased by the fact that the conversion to image format lies as completely outside of the control of document exploitation units as does the condition of original paper documents. In this instance, rather than specifying quality control for the scanning and import process, the most useful early processing would consist of image enhancement or recognition techniques that are automatically or semi-automatically tuned to the particular input documents, an area of active current research (e.g., Sarkar and Breuel, 2003; Veermachaneni and Nagy, 2007; Zavorin, et al., 2007)

In addition, the challenges of electronic document collection and exploitation are introduced here, such as identifying which image files in fact reflect documents or otherwise contain potentially recognizable text, and handling the massive volumes of data available electronically. When a system processes a full available set of electronic data, such as a hard drive or other general-purpose storage device, even identifying the type of processing appropriate for each file becomes non-trivial. File extensions provide a first indication of the likely type of data their files contain, but these may, of course, be misleading; internal file signatures offer a more certain method of recognizing

the data type, at the expense of additional processing time; further initial processing is required in order to determine the most suitable exploitation sequence among those applicable to the type of data. This is precisely the situation with data to be imported from D2C; among the image files, only a small percentage are likely to contain recognizable machine printed text, or handwriting suitable for the extent of recognition practicable, and identifying these files for OCR and similar processing in an efficient manner forms as much of a challenge as the recognition itself, and is equally important in order to target Machine Translation efforts effectively.

In this case, the challenges presented by volume and the unknown nature of the documents takes roughly equal place with the challenges stemming from the fact that the files represent only images of their textual content and interpretations of that content are required in order to produce the starting point for machine translation.

6 Electronic Text Documents

With textual documents acquired electronically, no OCR step is required; the text is directly present in the document. Therefore, these documents provide the most fertile ground for effective machine translation and other automatic processing of their content. This type of files forms the main bulk of the expected imports from the initial widespread deployment of the D2C system. (As the D2C system evolves, other file types may overtake these in volume, but the quantity and significance of these files will always remain very high.) In this case, the primary challenge comes from the large volume of electronic data and the heterogeneous and unknown nature of the collections. Issues that affect machine translation in this setting include those listed in the following subsections, organized by the type of triage they are most likely to affect. Note that most of these issues also apply to text images after OCR, but they are unlikely to appear salient in the presence of the much more sweeping effects of anything less than near-perfect OCR output.

6.1 Content-Based Triage

The Machine Translation factors that are most likely to affect content-based triage include any that are likely to impact the correct translation of im-

portant content-bearing individual words. Major elements include those listed below.

- The usual matters that affect machine translation on heterogeneous data apply here: use of dialect, writing that is so informal as to be difficult to parse or to process with standard linguistic knowledge and statistical models (e.g., chat, text messages, some e-mail), incorrect and creative spellings, etc. In this case, these matters take on importance only insofar as they alter significant words or complicate the choice of sense for those words; the effect on word order, function word choice and similar outcomes has far less importance in this case.
- Details of the extraction of plain text from various formats that encode this text can affect the input to machine translation and thus its results and effectiveness, for all purposes. For example, certain widely used tools for processing PDF files have only recently corrected an error that reversed Arabic words extracted from text-based PDF files; in addition, certain text-based PDF files may in fact encode sufficient information to display correct glyphs but insufficient information to map the internal representation of characters and their glyphs to standard underlying characters without introducing a process such as OCR. Even when all characters and words are extracted correctly, formatting information may be lost that would affect the meaning of these words and that is thus relevant to the translation process. For example, after extraction, a document may appear to contain a sentence that strings together several phrases with no direct relationships between them, because these phrases originally formed part of a bulleted list within the document.
- The significant factors in machine translation quality for this purpose may differ from those considered under other circumstances. For example, the fluency of the output plays no role in the use of MT output to support keyword-based retrieval, and note that it is also likely to have relatively minor significance in human review that focuses only on triage. Correct translation of the words that

indicate the primary subject matter, however, is clearly the most critical element for these purposes. Many other elements of quality begin to take on their importance at the point when human review occurs.

- Performance considerations become critical when handling massive quantities of data. A lower quality output produced quickly, and thus representing a full captured collection in near real-time, may offer far more value to the warfighter and local document exploitation team than an excellent quality output that operates slowly so that information about collections is not available within the same time frame as those collection operations. Similarly, translations that can run with lower memory and processing power requirements can be utilized by equipment available to a wider variety of personnel. This does not, of course, imply that more resource-intensive, higher-quality translation plays no appropriate role in intelligence operations; such translation becomes more important as the degree of human involvement and analysis increases, and even future automated triage may be able to take advantage of further cues in such full translations when they are available.

6.2 Metadata-Based Triage

In supporting the automated extraction of metadata, the question arises of whether machine translation should be performed before extraction of particular metadata values, which will then occur on the English approximate translation, or whether the extraction should be performed in the original language, and any necessary machine translation should be applied to the results. Note that this question is analogous to the decision in cross-language retrieval of whether to translate documents and perform queries on those translations, or to translate queries and retrieve documents in their original language.

Although in general extraction and interpretation is preferably performed within the original language, implying that translation should typically occur after this process, on its results, the most practical approach for each particular metadata value will depend on the specifics of extraction of

that value. Not only may there be occasional values that are more effectively extracted in translation, but other considerations may argue for applying extraction to translated text even when this is not the case. Some types of extraction will require software that is not available for all languages of interest; for at least those languages not addressed by available tools, using translation output will offer the only realistic short-term path to automatic identification of this data. Machine translation software typically runs most effectively on significantly sized blocks of running text that are grammatically well formed and provide surrounding context for the interpretation of individual words; the effects of applying machine translation to small isolated fragments of text identified as metadata may be more problematic than the loss of accuracy in initial extraction caused by operating on a translation, and this effect may be more difficult to mitigate for some types of metadata than for others.

For example, consider named entity extraction of people, organization, places, etc. mentioned in document text. The extraction process *per se* will prove more accurate when performed in language. Such extraction software is not available for all languages in which valuable content occurs, however. This fact argues for performing extraction on the translated text for unsupported languages. For languages with effective entity extraction tools, extraction within the original language remains preferable. This implies the use of different tools, which often use varying definitions of entity types (e.g., one tool may distinguish a facility from an organization or location, and another may categorize facilities as organizations when they stand in for them by metonymy, as in, “The White House announced today” and as locations when used directly, as in, “We met at the National Museum”); this difference presents a challenge but a surmountable one, since the relationships between these types can generally be identified and the number of such types remains relatively small. For entities extracted in the original language, machine translation should be performed to make them more useful. (Note that this refers to entity translation, not just transliteration; “Condoleezza Rice” is a much more useful name for a human reviewer than is “KwndwlyzA RAys”, and likewise, “American” is meaningful to an English speaker in a way that “al-Amirki” is not.) Entity translation is notoriously

problematic for machine translation systems; however, this effect can be mitigated by applying an entity-specific model for translation in this case. In recognition of this fact, the National Institute for Standards and Technology (NIST) included in its Advanced Content Exploitation (ACE) set of competitions for 2007 a track on entity translation, specifically, distinguishing between a “full” track of extracting and then translating entities from Arabic or Chinese into English, and a “diagnostic” track of performing the translations only, assuming perfect extraction.

A similar set of considerations and possibilities may lead to different conclusions and solutions for various types of metadata (document titles, copyright information, etc.).

6.3 Human Consideration Triage

Human consideration triages has much in common, in terms of its needs and priorities, with content-based triage. Most of the discussion in Subsection 6.1 also applies in this case, with several refinements.

- The typical machine translation issues of dialect, informal writing, and other challenges, as well as the issues of extracting text to translate from various formats, apply largely as described above. Their effects on the full details of the machine translation output take on greater importance here, however, as these may impact a human’s ability to determine the main subject matter and meaning of a document quickly and effectively.
- As mentioned above, the fluency of the output bears relatively minor significance in human review that focuses on triage. Correct translation of the words that indicate the primary subject matter remains the most critical element for these purposes. Many other elements of quality, however, such as preserving indications of emotion or tone, while secondary to this goal, may also prove very important to current or future intelligence requirements or to correct analysis of the data in combination with other sources. The degree of predictability of these needs remains an open question and a fruitful area for discussion within the community, as does the question of whether such needs can

be served more effectively by the production of explicit byproducts of machine translation representing determinations such as domain, register, and type of language, than by producing translations that attempt to preserve implicit linguistic indications of these phenomena.

- Performance considerations, while still of very high importance when handling massive quantities of data, become somewhat less critical in preparing data for human review, since the speed of the full process is typically limited primarily by the quantity and speed of human reviewers. When additional processing time or power can yield a significantly more accurate and informative translation, this application should be incorporated into a concept of operations as an additional step to the translation that provides immediate results, to support longer-term analysis and repeated triage decisions as intelligence requirements evolve. An area that requires more explicit determination of requirements is the appropriate tradeoff between speed and light footprints on the one hand, and accuracy and completeness of translations on the other.

7 Conclusions and Future Directions

The Harmony program's components illustrate the Government's needs for Machine Translation to aid in triage of collected documents, including both current needs that focus more on scanned paper and emerging needs that are expected to dominate in the near future and that focus on electronic collections made available through the D2C tool for cleaning the content of unsafe electronic material. As described above, these needs suggest several directions for Machine Translation that are the most likely to prove fruitful for these types of intelligence needs.

To support translation of the content of document images, whether scanned or collected electronically, working closely with OCR, handwriting recognition and similar technologies is called for, as is reflected in the structure and focus of the Government's MADCAT program. When working with electronic documents, it is necessary to account for the effects of text extraction, and the

more use that Machine Translation software can make of available structure and format information, the more fully informative its output will be. Specific attention is required for translating entities and other forms of automatically extracted, or extractable, metadata, both to support their identification and to enhance their use once they are extracted.

In general, a multi-tiered approach to translation, with several levels of detail and corresponding speed, from very fast and approximate, up through as precise a translation as possible, annotated with interpretations about tone and register, has the promise to support the range of needs required for intelligence triage and analysis. Such a system would allow for quick triage of vast collections, while maintaining the capacity for finer-grained processing of the most promising documents in the near future, as well as fuller processing of the entire collection to support later decisions in light of continuously emerging intelligence requirements.

References

- Dagan, I., Church, K., and Gale, W. A. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pp. 1-8, 1993.
- Sarkar, P. and Breuel, T. Amplifying accuracy through style consistency. In *Proceedings of the 2003 Symposium on Document Image Understanding Technology*, pp. 245-252, Greenbelt, MD, 2003."
- Taghva, K., Borsack, J., and Condit, A. Results of applying probabilistic IR to OCR Text. In *Research and Development in Information Retrieval*, pp. 204-211, 1994.
- Veerachaneni, S. and Nagy, G. Optimal interaction for style-constrained OCR. In *Procs. SPIE Symposium on Document Recognition and Retrieval*, Volume 6500, San Jose, CA, SPIE/IST, 2007.
- Zavorin, I., Borovikov, E., Borovikov, A., Hernandez, L., Summers, K., and Turner, M. A multi-evidence, multi-engine OCR system. In *Proceedings of the SPIE 19th Annual Symposium on Electronic Imaging Science and Technology, Document Recognition and Retrieval XIV*, San Jose, CA, 2007.