Parallel Corpus Development at NVTC

Jocelyn Phillips
Carol Van Ess-Dykema
National Virtual Translation Center
935 Pennsylvania Ave. NW LS-200
Washington, D.C. 20535
(202)962-9425
jocelyn.h.phillips@ugov.gov
carol.j.vaness-dykema@ugov.gov

Timothy Allison
Laurie Gerber
The MITRE Corporation
7515 Colshire Dr.
McLean, VA 22102
(703)983-2473
tallison@mitre.org
laurie.m.gerber@ugov.gov

ABSTRACT
In this paper, we describe the methods used to develop an exchangeable translation memory bank of sentence-aligned Mandarin Chinese - English sentences. This effort is part of a larger effort, initiated by the National Virtual Translation Center (NVTC), to foster collaboration and sharing of translation memory banks across the Intelligence Community and the Department of Defense. In this paper, we describe our corpus creation process – a largely automated process – highlighting the human interventions that are still deemed necessary. We conclude with a brief discussion of how this work will affect plans for NVTC's new translation management workflow and future research to increase the performance of the automated components of the corpus creation process.

Categories and Subject Descriptors
J.5 [**Arts and Humanities**]: Linguistics–*parallel corpus development and use.*

General Terms
Design, Standardization, Languages.

Keywords
Parallel corpora, translation memory, translation memory exchange (TMX), sentence alignment.

1. BACKGROUND
In its role as a leading center of translation technology for the Intelligence Community and the Department of Defense, the National Virtual Translation Center (NVTC) sponsors a multi-agency project whose participants collaborate on the development of exchangeable translation memory (TM) banks for government use. In early meetings, the participants decided on a preliminary goal of collecting and aligning Mandarin Chinese-English parallel documents in three specific domains. The goal of this parallel corpus development effort is to create TM for (operational) translation and experimental uses. Although we endeavored to automate the process as much as possible beginning with identifying sources of parallel documents and working through to the distribution of exchangeable TMs the process still requires human intervention [Koehn 2005]. This need for human review and correction will be present in many parallel corpora development processes.

2. PROCESS
2.1 Parallel Document Collection
Members of the participating agencies developed a corpus from legacy translated data. We identified a single source of extant parallel documents for collection and alignment – Open Source Center (OSC) translations of Mandarin Chinese journal articles[1]. This data source consists of original Mandarin Chinese documents, primarily text-based .pdf, with some image-based .pdf files, and their corresponding quality-reviewed .html English language translations.

We developed a file structure and naming convention, and created a processing log to keep the process consistent. This worked extremely well, in spite of our decision to retain intermediary files from each processing stage for future research purposes – a decision which rapidly increased the total number

[1] Since the Open Source Center produces these translations specifically for government use, and this project is designed to develop parallel corpora for government use, permission and copyright issues did not arise. Similarly, all Open Source Center translations are limited to government distribution, which is in line with project goals.

of files involved in our processing. We saved downloaded documents in a newly created folder whose name matched the unique identifier beginning each file name, e.g. a folder was created named Example1, and it included files Example1source.pdf (source-language .pdf) and Example1target.doc (target-language .doc).

The system was extended with subfolders and file name extensions for each process (text extraction, OCR, OCR cleanup, alignment, alignment cleanup). As files were processed, we moved the folder which held them (in case above, folder "Example1") through the file structure. High-level folders, which all files should pass through during processing, are:

- Raw Pairs
  - Needing OCR
  - Ready for Alignment
- OCRed Pairs
  - Needing Editing
  - Ready for Alignment
- Aligned Pairs
  - Needing Editing
  - Finished Alignments

When Example1 reaches its final processing destination (the "Finished Alignments" subfolder), the folder will hold any intermediate files created (for example, unedited OCR output Example1source_uneditedOCR.doc), in addition to the original files Example1source.pdf and Example1target.doc. Future research may explore the possibility of using the intermediate files to train models to automate processes that must currently be undertaken manually, such as OCR correction.

For every file that went through this process, the processing log recorded file name, file location, file type, file size (words), file size (KB), source language file format (image-based .pdf, or text-based .pdfs), domain, and genre. We manually noted domain and genre in the processing log at the time of document capture, in accordance with NVTC standardized domain and genre lists[2].

[2] NVTC uses a domain list obtained from the Joint International Annual Meeting on Computer-Assisted Translation and Terminology (JIAMCATT) 2008. NVTC uses a genre list adapted from the NVTC DHDS-T (Deployed Harmony DOCEX System – Template) Appendix B "Definition of Document Types".

2.2 Optical Character Recognition

The NVTC Translation Technology Assessment Laboratory currently has two OCR systems installed – referred to hereafter as System 1 and System 2. Both of these systems are capable of Chinese language OCR, and we tested both of them for the purpose of converting the OSC image files into text.

Our preliminary manual inspection of the OCR output of both of these systems showed that the systems' error rates were low enough (<5% of total output document characters misidentified) to be acceptable for the project's needs. Given this low error rate, we decided that OCR correction would not be performed on the majority of the documents. However, System 1 provided full dictionary support for Mandarin Chinese; that is, upon "reading" the document, System 1 provided an editable draft copy of the scanned text, with likely errors highlighted, and a list of potential replacement characters provided. This capability facilitated in-program error correction on the rare occasion that errors were egregious enough to necessitate correction. Therefore, we chose System 1 as the OCR tool for this project.

Because the image quality of the documents needing OCR for this project was relatively high, many image analysis steps in the OCR process could be automated. Thus, before opening any image files in the OCR system, we created a template in order to automate these steps – including identifying the ingest language of the images files and performing automatic skew correction, page orientation correction, and page analysis. Page analysis, often referred to as "zoning," identifies areas of the document that are to be recognized by the OCR system. Once the template was established, the OCR system automatically analyzed files as we opened them.

Following automatic ingest, a linguist performed a quick scan of the OCR analysis to determine if additional review is necessary. A two-monitor setup presents information for a very ergonomic process:

1. Open target translation document for reference (on second monitor screen in a two-monitor set-up.)

2. Manually review automatic "zoning."
    a. De-zone material with no counterpart in the target language document (e.g., images, page numbers.)
    b. Confirm zones were assigned the correct sequence in the text.
3. Have OCR system perform recognition and produce .doc output
4. Scan output for major errors.

5a. If output requires editing, folder moved to "OCRed Pairs>Needing Editing" folder. Perform editing in the OCR interface.

5b. When output is acceptable, move folder (also containing intermediate processing files) to "OCRed Pairs>Ready For Alignment" folder.

2.3 Alignment

Once all files were in text format, we aligned the files using stand-alone bilingual text alignment software. First, we created an alignment project template, which would automate many steps of the alignment process, such as source and target language identification, segmentation choices (e.g., whether or not to segment at a semi-colon) and identification of the desired output format.

We aligned documents on an individual basis, rather than batch aligning them all as part of one large alignment project, in order to create .tmx files that corresponded to original documents. We did this in order to ensure that all translation memory translation units could be marked with domain, genre, and security classification metadata derived from the characteristics of the source document. Aligning individual files also allowed for easier human review and editing, due to the difficulty of visually scanning an alignment composed of tens of thousands of individual translation units.

Originally, we had assumed that text-based .pdfs would be easier to work with than image files, due to the fact that text-based .pdfs did not need to go through the extra step of image conversion through OCR. However, when text-based .pdf source documents were directly aligned with their target language .doc counterparts, the resultant alignments were often laden with errors due to formatting issues with the text-based .pdf format. Despite the fact that the text in text-based .pdfs should be accessible with no further processing, and the fact that OCR will inevitably result in a document with some misrecognized characters, the OCR process yielded cleaner alignments than text-based .pdf files did. Thus, we found that the OCR process (when combined with manual review of page analysis) often produced a file that was easier to align than text-based .pdf extraction.

Due to formatting issues with text-based .pdfs and the fact that the aligner had no language-specific processing tools, all created alignments needed to be manually reviewed and edited. Alignments created from OCR-processed files generated different types of errors than those created from text-based .pdfs; these different types of errors were easily visible in the manual alignment review stage:

For alignments following an OCR process:

- Generally, minimal editing was necessary.
- Most common alignment error: OCR system had identified a speck on the original image file as a punctuation mark, and consequently the aligner had segmented at that mark.

For alignments from text-based .pdf files, the alignment software automatically extracted the text portion. However, this process is error prone:

- Generally, more serious alignment errors were prevalent.
- Errors varied widely, but included:
    - entire sections of source-language text being omitted from the alignment.
    - source document page metadata (header, footer, page number, authorship information) being included and incorrectly aligned with target document text.
    - source document segmentation occurring at line breaks (instead of at punctuation marks).

2.4 Metadata Markup and Document Aggregation

Once the alignments had been created and edited, we had effectively developed an aligned parallel corpus from parallel documents. For this project, the corpus was in the form of a series of .tmx files. Before aggregating these .tmx files into a single TM,

metadata was to be added to the .tmx files, such that the domain, genre, and security classification level of every translation unit could be identified. These metadata derived from characteristics of the source documents from which the alignments had been made. Since, at this point in the processing, each .tmx file corresponded to only one original source-target document pair, the metadata could be applied to all translation units in each individual .tmx file. The metadata was inserted into each .tmx file in the process of aggregating the .tmx files into one large TM, according to the following process:

1. A new, empty TM was created in a TM system
2. Mandatory domain, genre, and security classification fields were defined within the TM system for this TM
3. Picklists were created for each field, to standardize and limit the labels available for each field.
4. Individual .tmx files were imported into the new, empty TM
5. Upon ingest of these individual files, the TM system prompted for field labels to be selected from the picklist.
6. Selecting picklist values automatically generated XML property elements for each translation unit identifying domain, genre, and security classification.

It is worth noting that, in line with the experimental TM goals of this project, portability experiments were undertaken, wherein the aggregated TM with its translation unit-specific metadata was imported and exported to and from different TM systems. All of these systems were able to maintain all language data in the .tmx upon ingest, but many of the systems could not manage the importation of the metadata. As these systems develop to allow for import and export of metadata, best practices for incorporating translation-unit specific metadata into .tmx will be worth exploring.

2.5 Corpus Creation Productivity

We did not originally set a formal goal for corpus size upon beginning the corpus creation effort. The corpus contained 15,000 translation units as a result of five months of non-continuous work. When our creation process had been fully established, a linguist with experience with the system could create 250-300 translation units per hour. This estimate takes into account every part of the corpus creation process – from identifying files to aggregating the .tmx files into one TM.

Once the corpus creation process was established, we taught a linguist completely new to the field of parallel corpora to create translation units using the system. On average, the new linguist could create 80-120 translation units per hour during the first week using the system; by the second week using the system, the new linguist's creation rate had increased to 100-200 translation units per hour.

In both cases, the linguists were skilled in Chinese-English translation. However, since the aligner uses no language processing, and instead makes segmentation and alignment decisions based largely on punctuation, professionals with less language ability could also perform corpus creation using this process. The ability of the professional with no knowledge of the source language to perform alignment correction would be dependent on the nature of the source language text; source language text with more identifiable non-textual markers (numbers, formatting, etc.) will be more easily parsed by a professional unfamiliar with the source language.

Another impediment to the corpus development process for a professional with no source language capability would be the lack of ability to correct (and in some cases even identify) OCR errors. OCR correction was a minimal part of our corpus creation process, but it would be a factor to consider when document image quality is low and the corpus creation goal is to create high-quality, usable alignments.

3. EXPERIMENTAL USAGE

The product resulting from this parallel corpus development process was a TM that was to be shared with the project participants. Because the participants will be using a variety of TM software packages, we conducted a series of experiments in order to discover any issues that might affect the integrity of the TM as it was used in and shared between different TM systems.

When we began assessing the usability of our TM on different TM systems, our TM consisted of approximately 15,000 segments. Understanding that

corpus size has a significant impact on TM performance, we have begun tests to assess corpus size thresholds for usefulness in translation projects. We will report on all available reports at the conference. The presentation itself should be available on the conference website and MT Archive.

In order to assess TM portability, we imported and exported the TM created from our newly developed parallel corpus between three different TM systems. We soon discovered a corpus development-related challenge to TM usability and portability. We discovered that segmentation choices made by the linguist during manual alignment correction sometimes differed from segmentation choices made by the different TM systems as new documents were parsed for translation. These differences in segmentation represented a serious impediment to the TM system's ability to find matches.

4. SUMMARY

Initially, we anticipated that the real findings in this project would come from experiments on TM in an (operational) translation setting. However, to date, the corpus building effort as summarized above has generated significant findings independent of the operational usage of TM created from this corpus.

First, the extraction of text from documents was not as straightforward of a process as initially expected. Text-based .pdfs, which by their very nature should allow for relatively straightforward text extraction, can present a variety of challenges to systems (such as aligners) which call for text inputs.

Second, given the lack of language processing in the alignment software used for this project, challenges can arise when documents to be aligned to one another are in different formats. Without language processing capabilities, the aligner is not capable of recognizing, for instance, that the header on a source-language .pdf is not translated in a .doc target-language file, and this difference may cause alignment error. Essentially, if an aligner is devoid of language processing capabilities, any differences between source and target language documents (besides language) may present challenges to the alignment process [Gale and Church 1991].

Given that the project's intention was to create usable TM from the developed corpus, the alignments created needed to be as high-quality as possible. In order to create high-quality alignments, and in light of the hindrances to automated high-quality alignment described above, we found that a manual review and editing stage was essential in the alignment process. Review and editing was also necessary in the OCR stage, but generally far less editing was required at the OCR stage than at the alignment stage, in part because many of the alignment issues stemming from text-based .pdfs can be corrected for in the automated OCR process.

We also found that, within the alignment editing stage, it is necessary to ensure that the linguist performing the manual editing is aware of the impact that manual segmentation decisions may make on a TM system's ability to perform matching against a document segmented differently.

5. FUTURE WORK

In addition to continuing to develop shareable sentence-aligned corpora among project agencies, there are two avenues at NVTC that the work presented in this paper will inform: the new NVTC translation management workflow and our applied research program.

5.1 New NVTC Translation Management Workflow

As part of a larger ongoing effort to improve the translation workflow at NVTC [Van Ess-Dykema et al. 2009], we have developed the workflow depicted in Figure 1. The goal of this workflow is to leverage parallel corpora to enable our translators' work to feed back into machine translation (MT), TM and terminology management tools. The work in this paper makes clear the need for several functions in our planned workflow. A paralinguist or team of paralinguists will prepare the incoming documents for ease of ingest into a translation memory system. The paralinguist will also coordinate with the task manager to select appropriate translators and memory banks for incoming tasks. The corpus quality control expert will assure that aligned sentences generated in our workflow are accurately aligned and contain the correct metadata. A parallel corpus engineer/researcher will be responsible for the sentence-aligning of legacy parallel documents, and advising on and carrying out applied research to support the workflow and translators. NVTC will

include the metadata on our parallel corpora in its entries in the Intelligence Community Lexical Resources (ICLR) catalog,

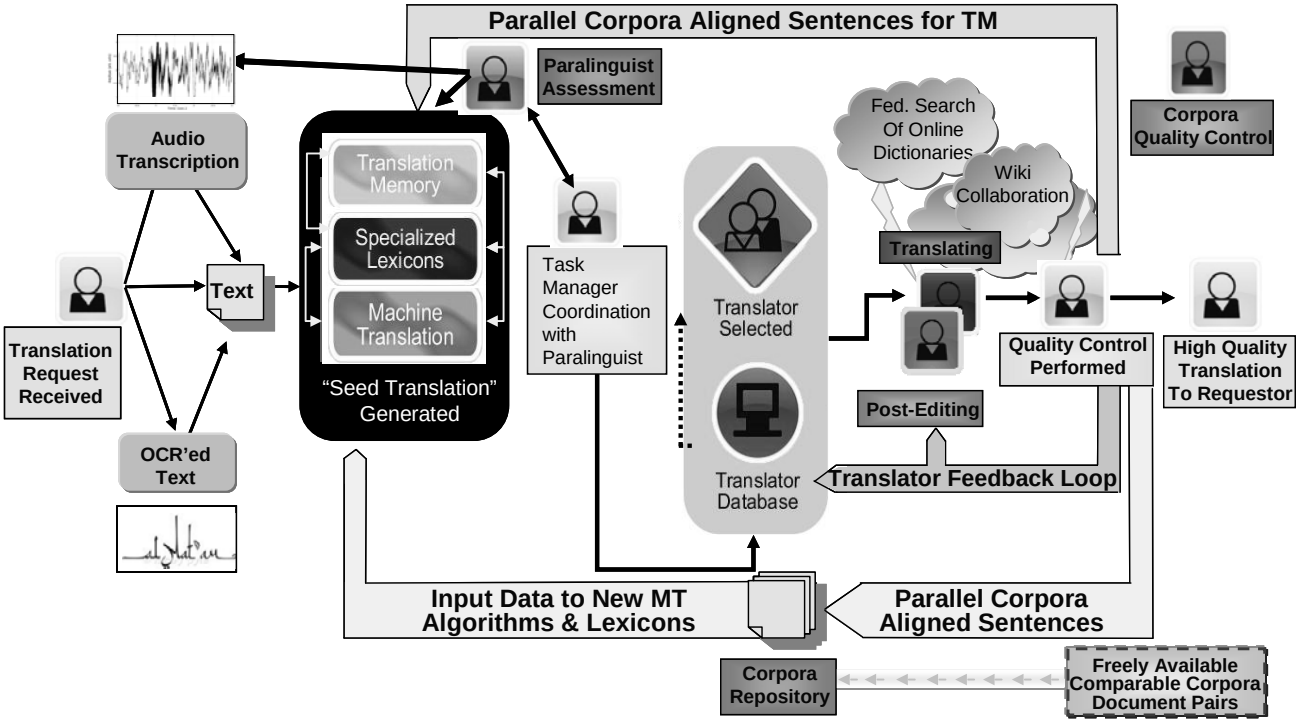

Figure 1. NVTC New Translation Management Workflow

5.2 Research Directions

Given our findings from this initial effort, we will focus our research efforts in five areas:

- Alignment tools. We hypothesize that alignment tools that use a translation model will require less post-editing than did the tool we used [Church 1993; Melamed 1999; Moore 2002].
- Pdf extraction tools and methods. We need to delineate the different types of challenges that can arise and offer methods to reduce the amount of required human intervention.
- For text-based .pdfs, we will apply techniques from the natural language processing community to join appropriate blocks of text and strip headers.
- The effects of alignment errors on downstream processing, whether that be for MT or TM. Our alignment process required much human intervention to yield high quality alignments. We need to determine whether that level of quality is necessary for both MT and TM.
- The nature of our legacy corpora. We will calculate the repetitiveness of our corpora to give insight into the number of new segments that would require no translation or only slight edits. In turn, this will allow us to estimate savings for given genres or language pairs.
- The characteristics of corpora necessary for successful use in TM tools. How large of a corpus of parallel sentences, how much repetition and what types of repetition yield useful results for translators given the nature of NVTC translation tasks?

6. REFERENCES

[1] CHURCH, K. 1993. “Char_align: a program for aligning parallel texts.” In *Proceedings of Association for Computational Linguistics 1993*. Columbus, OH, 1993.

[2] GALE, W. AND CHURCH, K. 1991. “A Program for Aligning Sentences in Bilingual Corpora.” In *Proceedings of Association for Computational Linguistics 1991*. Berkeley, CA, 1991.

[3] KOEHN, P. 2005. “Europarl: A Parallel Corpus for Statistical Machine Translation.” In *Proceedings of the Tenth Machine Translation Summit 2005.* Phuket, Thailand, 2005.

[4] MELAMED, D. 1999. “Bitext maps and alignment via pattern recognition.” *Computational Linguistics, 25(1),* 107-130.

[5] MOORE, R. 2002. “Fast and accurate sentence alignment of bilingual corpora.” In *Proceedings of the 5th Biennial Conference of the Association for Machine Translation in the Americas*. Tiburon, CA, 2002.

[6] VAN-ESS DYKEMA, C., PERZANOWSKI, D., WHITE, J., AND CONVERSE, S. 2009c. “Exploring Translation Memory for Extensibility Across Genres: Implications for Usage and Metrics.” In *Translating and the Computer 31. Proceedings of the Conference*. London, UK, 2009.