MTriage: Web-enabled Software for the Creation, Machine Translation, and Annotation of Smart Documents

Reginald Hobbs
Multilingual Computing Branch
Army Research Lab
Adelphi, MD
hobbs@arl.army.mil

Jamal Laoudi
Advanced Resource Technology, Inc.
1555 King Street Suite 400
Alexandria, VA
jlaoudi@arl.army.mil

Clare R. Voss
Multilingual Computing Branch
Army Research Lab
Adelphi, MD
voss@arl.army.mil

Abstract

Progress in the Machine Translation (MT) research community, particularly for statistical approaches, is intensely data-driven. Acquiring source language documents for testing, creating training datasets for customized MT lexicons, and building parallel corpora for MT evaluation require translators and non-native speaking analysts to handle large document collections. These collections are further complicated by differences in format, encoding, source media, and access to metadata describing the documents. Automated tools that allow language professionals to quickly annotate, translate, and evaluate foreign language documents are essential to improving MT quality and efficacy. The purpose of this paper is present our research approach to improving MT through pre-processing source language documents. In particular, we will discuss the development and use of MTriage, an application environment that enables the translator to markup documents with metadata for MT parameterization and routing. The use of MTriage as a web-enabled front end to multiple MT engines has leveraged the capabilities of our human translators for creating lexicons from NFW (Not-Found-Word) lists, writing reference translations, and creating parallel corpora for MT development and evaluation.

1. Introduction

Our organizational focus is basic and applied research in automated machine translation of foreign language documents, supplying our users with techniques and criteria for the use of MT. We have designed embedded MT systems, such as FALCon (Forward Area Language Converter), to assist the non-linguist in handling foreign language texts. (Fisher & Voss, 1997) The problem space that we are trying to address is: given multiple available MT engines and a lack of sufficient translators, how can we leverage human expertise to handle large collections of foreign language documents? Our answer is to identify techniques and design tools to capture the annotations from the analyst, using the information for selection and routing to MT services.

2. Approach

One aspect of HLT (Human Language Technology) research here at ARL involves how best to evaluate available or constructed MT engines against user needs. To keep up-to-date with the status of evaluation in the MT community, we have been involved in the annual NIST (National Institute of Standards and Technology) Evaluation process and workshops. During a NIST Evaluation, participants are supplied with a large dataset in a selected foreign language. These datasets are to be translated using available MT engines to create output documents for scoring through automated metrics. Each organization attempts to maximize scores through configuration of their MT engines.

One of the more labor intensive parts of our participation in the NIST Evaluation was analyzing the datasets. (Voss et al, 2006) The primary contextual information supplied with the documents consisted of several designated genres for binning purposes. Our translator created reference translations of samples from the dataset to identify other potentially helpful information. Determining configuration parameters for MT was an iterative, trial-and-error process. This was particularly the case when multiple MT engines were available; because the training corpora for our statistical MT servers were necessarily different than the NIST input, it required sometimes visual inspections of multiple MT outputs to detect sensitivity to changes in parameters or input data. It was clear that we needed a way to automate some of the data analysis process, leveraging the time and expertise of our translator to handle large collections quickly.

Our experience with the NIST Evaluation was a catalyst for our pre-processing for MT (or pre-MT) approach. We had some success with enhancing MT by leveraging linguistic and contextual knowledge to route documents to the appropriate MT engines. The hypothesis of our pre-MT research is that there are characteristics of foreign language texts that can be exploited to improve the output of MT. This concept uses the software engineering metaphor of an MT engine as a "black box", focusing on improving what comes out of the box by enhancing what is fed into it. Our research method is to iterate through: 1) identifying candidate pre-MT activities, 2) using tools to gather, annotate, and construct test datasets for each pre-MT, 3) defining MT routing and configuration techniques from the annotations, 4) translating using multiple engines, and 5) evaluating the output with automated metrics.

Our research focus for this year's NIST Evaluation was to 1) leverage the expertise of our translators for analyzing NFW and named entity translation for augmenting MT with new lexicons, and 2) take advantage of MTriage as a general purpose tool for accessing multiple MT engines as web services for creating parallel text.

3. Tool Development

We required a tool to support pre-MT activities that could serve as a front-end to multiple MT engines as well as a mechanism for capturing and recording annotations from the translator. The desirable features for the tool: 1) Web-enabled, service oriented architecture (SOA) tool for platform-independent access to multiple MT engines, 2) capabilities for editing/storing annotations of source documents, and 3) support for the creation of reference translations.

3.1 MTriage (Machine Translation Input Annotation and Genre Encoding)

The MTriage prototype is implemented as a JavaScript client. The choice of JavaScript allowed for rapid, iterative development as well as a measure of platform independence. The tool could be utilized in any standard web browser. Distribution and installation of the latest version of the tool was very straightforward because the web file containing the embedded MTriage client could be easily e-mailed to the user for testing and analysis.

The other major design decision for creating the tool as a web-client was the necessity for remote access to several available MT engines. Each engine was set up as a remote web service, accessible by SOAP (Simple Object Access Protocol) messages across the network. The format and method signature of calls to the remote MT services were described in WSDL (Web Services Definition Language) files supplied to developer. Dynamic configuration of the MT engines was made available to the user directly through the user interface.

Location 1 (denoted by the numbered circle) on the screenshot of MTriage in Figure 2 on page 5, shows the location of the text input area for the source document. The language mapping pull-down menu above the input area allows for the selection of language pairs for translation. In the example in the figure, Arabic-to-English translation will be performed on the Arabic script from the source input area.

Location 2 is a grouping of some of the annotations that can be associated with the text. The type of file to be processed (local file, online document, or generated text) can be selected as well as a lexicon by choosing an appropriate domain. The domain selection controls dictionary stacking within the called MT engine. The named entity option uses pattern matching to detect hits from a pre-designated list of unique names. These identified named entities, their location, and appropriate translations are annotated.

The analyst can choose from several MT engines (as shown at *Location 3*). The selection of the MT engines is color-coded to indicate their availability. The color coding is changed upon selection of a language pair. *Green* indicates that an MT engine supports the language pair, *red* indicates that the MT engine cannot process the language pair, and *blue* denotes the currently selected MT engine. A user can enter the file to be translated directly in the *File:* text field or can browse the local directory structure to select a file for uploading into the text area.

Location 4 is the text area that contains the results of the automated translations. Human translations are entered directly in this text area to create reference translations for evaluation. The resulting reference translation is annotated with the translator's name and explicitly connected with the source document. Post-edits of MT output can also be directly entered into the target results area, and are denoted as such within the internal smart document.

The translation controls and configuration options are depicted at *Location 5*. The options are also color-coded and are MT-dependent. Upon selecting an MT engine, the configuration options are color coded appropriately to green , blue, or red. The user of the MTriage tool can toggle transliteration on the output data, as well as request normalization of the source data by selecting these options. An algorithm within the tool for calculating the number of NFWs is used to display the metric directly on the user interface after translation. The algorithm is MT-specific, based on the presence of delimiters or other notations that distinguish the NFWs from the rest of the text. Translating a document with the default settings is done by pressing the *Translate* button. Using the *TranslateWithOptions* button causes MTriage to parameterize the configuration options selected by the user.

Viewing the internal smart document is done by right-clicking on the user interface and choosing that option from a pop-up context menu. Exporting the internal smart document to multiple parallel text files, creating a multilingual Excel spreadsheet, and saving the smart document in XML format are functions that are also accessible on the context menu.

3.2 Smart Document Schema

The annotations had to be associated with the source document in a way that did not modify the original content. Incorporating markup directly within the text would require additional processing to strip out the details prior to translation. The smart document approach treats the document as an object, encapsulating the raw input text within metadata and annotations that describe it. Unlike the

Annotate tool, which uses linguistic information and hidden Markov models for syntactic corpus annotations (Brants & Plaehn, 2000), we required a static structure for capturing annotations within the documents.

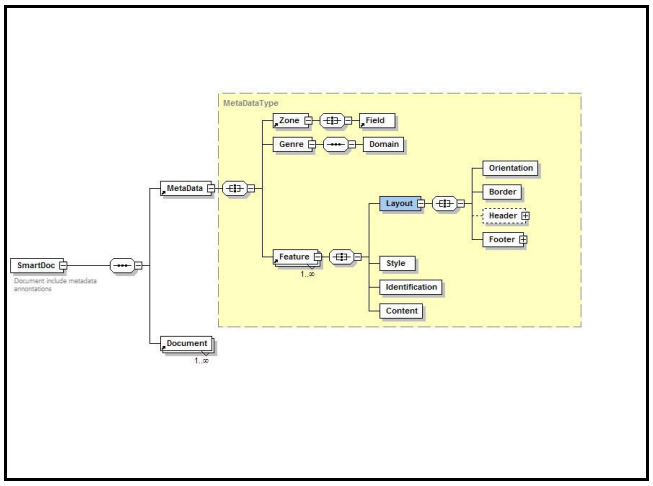

Figure 1 Smart Document Schema

The choice of an XML schema supports the object-oriented view of the document, separating the semantically useful smart document structure from the applications that take advantage of the metadata. Annotea, a RDF standard developed by the worldwide web consortium (W3C), supports web based document sharing and collaboration through markup tags, bookmarks, and annotations. (Kahan et al, 2001) An XML schema, called WandaML, has been developed to capture single language document annotations for the purpose of forensic handwriting analysis. (Franke, Guyon, & Schomaker, 2004) In a similar fashion, our smart document schema takes advantage of the extensibility and flexibility of a markup language for annotations. It differs in that we are not constraining our documents with respect to genre (web documents), monolingual text, or specific application (handwriting analysis). The use of XML namespaces will allow us to leverage the capabilities in these other languages, focusing our efforts on capturing annotations for pre-processing activities.

Figure 1 is a visual representation of the smart document XML schema. The important thing to note about the design of this knowledge structure is that it describes a smart document as containing multiple documents and the metadata about those documents. The expanded area of the schema (highlighted in yellow) shows some of the attributes about a source document that could be attached, such as the locations of zones and fields within the text, the genre of the document, layout features, the presence of mixed language texts, and so on. These attribute annotations can be assigned to the document computationally, through user input, or a combination of both. We plan to examine the use of machine learning for defining smart documents within MTriage, potentially creating feature vectors to automate annotations for large datasets. One of the primary advantages of XML as knowledge format for smart documents is the ability to use XSL for transformations of the documents, supporting multiple MT applications independent of the original source of the material. For now, MTriage constructs the internal smart documents, but uses embedded functions to generate the parallel texts, spreadsheets, and network messages for the MT web services.

4. Building Corpora with MTriage

The MTriage tool is being used to build the datasets necessary to evaluate the pre-MT approach. In Figure 3 on page 6, a smart document containing Arabic source text is encapsulated with translated English results documents. The metadata for this particular document contains tokenization information that was generated by MTriage. Between the <*tokens*> tags, there are ordered pairs of numbers representing the location and string length (respectively) of all the tokens in the Urdu source text. Using the Unicode representation of the tokens, it is possible to separate the tokens into sub-types. The <*tokens*> tag with the *punctuation* type attribute contains order pairs where the first value is the detected punctuation character and the second value is its position in the original source text. Using the same method, MTriage can identify number tokens and potential NFWs as tokens that are using an alternate script that the primary text. The automated script identification of "Arabic" is based on the Unicode range.

There is an automated translation from *MTEngine5* within the smart document. The NFW count for this particular engine can be determined, since the bracket "[]" delimiter is used around the transliterated text. The reference document created by our translator is included as a ground truth English version of the Arabic text. The *variant* attribute on both non-source documents identify them as translations and include language pair mapping information. According to our translator, the capability for visual inspections of parallel MT outputs alone has enhanced the data gathering and analysis process.

A case study where the usefulness of MTriage was recently validated was during ARL's participation in the Urdu Language Track for the NIST Evaluation. We focused on research issues of augmenting MT engines that support Urdu-to-English translation. (Voss et al, 2008) Urdu, as a Less Commonly Taught Language (LTCL), had fewer resources available in terms of both MT engines and parallel corpora for building engines. Because MT engines are fairly static with respect to lexicon, translators encounter increasing numbers of NFWs when trying to process large datasets. From an MT stakeholder

perspective, there is the question of how much data is necessary for improving their existing engines and how to effectively apply the data. To answer these questions, we conducted empirical studies with the NIST evaluation dataset.

The MT engines that support the Urdu-to-English language pairs were accessible through MTriage. The translators and analysts were able to use the tool for translating portions of the NIST datasets and specific data gathering tasks. One particular use of MTriage was in harvesting NFWs to create lexicons for MT engines. Source language text was collected from the NIST Urdu resources as well as through a Web scraping application. The Web scraper analyzed Urdu news sites, extracting the Urdu text on a daily basis. Our Urdu translator could then use MTriage to translate the source text against the Urdu MT engines to identify words and phrases that were not part of the engine's current lexicon. Figure 4 on page 7 is an Excel spreadsheet generated by MTriage that shows the MT output as parallel text. MTriage used the internal smart document structure to generate a spreadsheet with the original source language text and two translations using *MTEngine5*. MTriage used metadata about the source document to set the appropriate text direction, in this instance, right-to-left. In the second column on the sheet, the translation shows NFWs as transliterated Latin characters within brackets. To generate the third column, the translator turn off the transliteration option, leaving the Urdu string in the text. Giving the translator the ability to examine the new word in its original context helps with word sense disambiguation and supports a more accurate reference translation for the new lexicon.

We constructed several lexicons using this technique for use in our augmentation experiments. For one of the MT engines, we converted the lexicons to customized user dictionaries for stacking within the engine. We then ran a series of experiments of the NIST evaluation dataset, using MTriage to parameterize the selection of user dictionary. We evaluated the MT output using the assessed BLEU score against the baseline (no custom Urdu dictionary). There was a slight increase of BLEU score in one case, but it was too small an increase to be statistically significant. The other scores remained about the same as the baseline, or decreased by a small amount. The lack of major improvement could be due to the sample size or a lack of sensitivity to this type of augmentation within the BLEU automated metric. The NFW count for the evaluation data did improve, however, with up to ten percent of the new words being detected by the MT engine. It would be interesting to see if there is a corresponding increase in the measure of effectiveness of the MT output, i.e. the usability of the MT output for specific tasks.

5. Future Work

The potential improvement on MT quality through annotation-driven pre-processing activities seems intuitive and supported anecdotally by our efforts so far. We are using the MTriage tool and smart document schema in an empirical study with automated metrics to validate this improvement statistically. The smart document format is being leveraged to support other MT data requirements, such as building parallel corpora for training datasets, segmentation and alignment of source documents, and comparison of multiple human translations to calculate inter-annotator agreement. Incorporating adaptive machine learning techniques, as described in the GATE (General Architecture for Text Engineering) (Cunningham et al, 2002) environment, into MTriage is being examined for automated annotation and information extraction on large data collections.

6. References

Brants, T. and O. Plaehn. (2000) Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. (LREC '00). Athens, Greece. May 31 - June 2.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan. (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02). Philadelphia, July 2002.

Fisher, F. and C.R. Voss (1997) FALCon, An MT System Support Tool for Non-linguists. In *Proceedings of the Advanced Information Processing and Analysis Conference* (AIPA 97), McLean, VA.

Franke, K., L. Guyon, L. Schomaker (2004). The WandaML markup language for digital document annotation *In Proceedings of 9th IWFHR*, Japan, Los Alamitos: IEEE Computer Society, pp. 563-568.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., and Swick, R. R. (2001) Annotea: An Open RDF Infrastructure for Shared Web Annotations. In *The Tenth Int'l World Wide Web Conference*, Hong Kong, May, pp. 623-632

Voss, C.R, J. Laoudi, S. Lee, J. Micher, C. Tate. (2006) Five Arabic-English MT Engines: Performance on Newswire & Military Corpora. *In Proceedings of the DARPA TIDES/NIST Workshop on Machine Translation,* Alexandria, VA.

Voss, C., R. Hobbs, J. Micher, J. Laoudi, M. Aguirre, R. Chang, G. Jahed. (2008) Augmenting Urdu-English MT Engines at ARL. Presentation, In *Proceedings of the NIST Machine Translation Workshop*, Arlington, VA.

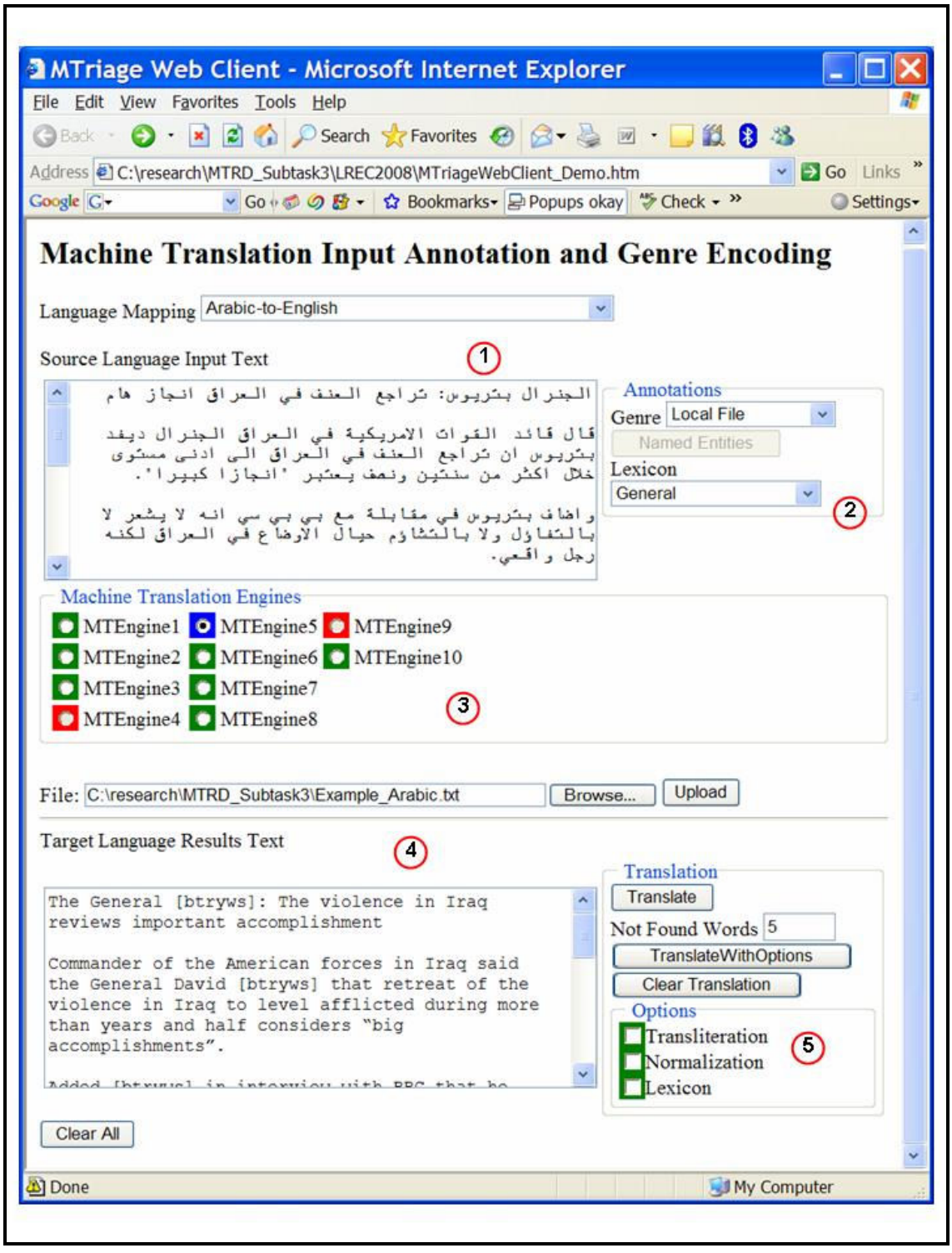

Figure 2 MTriage User Interface

```
<?xml version="1.0" encoding="utf-8" ?>
- <smartDocument xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema" xsi:noNamespaceSchemaLocation="SmartDocument.xsd">
  - <meta-data>
      <tokens>(0,3), (4,4),(9,6),(16,8),(25,2),(28,5),(34,4), (39,5),(45,2),(48,4),(53,5),(59,4),(64,2), (67,5),(73,4),(78,4),
        (75,2),(86,5),(92,3), (96,5),(104,2),(107,3),(111,12),(124,10),(135,2), (138,5),(144,4),(149,6),(156,2),(159,2),
        (166,1), (162,4),(170,4),(175,5),(181,12),(194,6),(201,4), (206,4),(211,5),(217,4),(222,3),(226,4),(231,6),
        (238,2),(241,2),(244,5),(250,6),(257,3),(261,4), (266,12),(279,6),(286,8),(295,2),(298,6),(305,7), (322,1),(313,9)
        </tokens>
      <tokens type="punctuation">(.|166),(.|322)</tokens>
    </meta-data>
  - <document uri="http://www.aljazeera.net/NR/exeres/85D4A217-155D-4CC7-B84F-1B004FE4D548.htm" genre="online"
      domain="General" language="Arabic" script="Arabic" creation="Tuesday, April 01, 2008 4:35:19 PM" characters="325"
      numberOfTokens="57" variant="original">
      <content>وقد دافع السفير الأميركي في بغداد ريان كروكر عن قرار رايس، وأكد في مؤتمر صحفي عقده في إمارة دبي اليوم أن على الدبلوماسيين الأميركيين أن يلبوا خدمة بلادهم في أي مكان.
        وذكر كروكر الدبلوماسيون بالقسم الذي أدوه والذي يشدد على خدمة أميركا في أي مكان، مساويا بين مهمة الدبلوماسيين وأولئك العاملين في القوات المسلحة الأميركية.</content>
    </document>
  - <document uri="Reference Translation" language="English" mapping="Arabic-to-English" variant="translation"
      author="Jamal Laoudi" creation="Tuesday, April 01, 2008 4:38:22 PM" characters="394" words="59">
      <content>Ryan Crocker, American ambassador in Baghdad defended Rice's decision, emphasizing in a press
        conference held today in Dubai emirate that American diplomats must answer the call to serve their country
        anywhere. Croker reminded the diplomats of the oath the took and that emphasized serving America anywhere,
        liking the diplomats' mission to that of those serving in the American armed forces</content>
    </document>
  - <document uri="Generated By MTEngine5" language="English" mapping="Arabic-to-English" variant="translation"
      author="MTEngine5" creation="Tuesday, April 01, 2008 4:37:22 PM" characters="444" words="69" evaluation="nfws:3">
      <content>The American ambassador in Baghdad defended juicy [krwkr] about decision of Rice, and press conference
        confirms in complicates him in bearish Emirate today That on the American diplomats that [ylbwaa] their service of
        countries in any place. Male [krwkr] the diplomats in the divisions who afflicted him and who service of America in
        any place intensifies on, strickles between task diplomatic and those laborers in the armed forces
        American.</content>
    </document>
</smartDocument>
```

Figure 3 Example Smart Document with Annotations

Source Language Text	***MTEngine5 MT Output***	***MTEngine5 MT Output***
Urdu	***English***	***English***
’اتر پردیش میں آئی ایس آئی کا پھیلاؤ‘	Expansion `of ISI in Uttar Pradesh '	Expansion `of ISI in Uttar Pradesh '
واضح رہے کہ حالیہ دنوں میں ریاستی دارالحکومت لکھنؤ اور بجنور ضلع سے پولیس نے پانچ مبینہ شدت پسندوں کو گرفتار کیا ہے جن پر ممنوعہ شدت پسند تنظیم حرکت الجہادالاسلامی سے وابستہ ہونے کا الزام ہے۔ اس کے علاوہ تین انتہا پسند کولکتہ سے گرفتار کرکے لکھنؤ لائےگئے ہیں۔ جلال الدین عرف بابو، نوشاد، عزیز الرحمان، مختار اور اکبر نام کے یہ مبینہ انتہا پسند فی الوقت لکھنؤ کی خصوصی سی بی آئی عدالت کی ہدایت پر پولیس ریمانڈ پر ہیں۔	Remember that police from state capital Lucknow and [bjnor] district in recent days has arrested to five alleged extremists on which prohibited extremist is allegation of being related with organization movement [aljhadalaslami]. Of arresting with three extremists Calcutta besides him are Lucknow [la-eeg-ee]. Mukhtar and Akbar Aziz Al Rahman, Naushad, known as gentleman, Jalaluddin is on police remand on the instructions of court special C B I of Lucknow at present this alleged extremist of name.	Remember that police from state capital Lucknow and بجنـــور district in recent days has arrested to five alleged extremists on which prohibited extremist is allegation of being related with organization movement ىالجہادالاســـلام. Of arresting with three extremists Calcutta besides him are Lucknow لائےگئـــــے. Mukhtar and Akbar Aziz Al Rahman, Naushad, known as gentleman, Jalaluddin is on police remand on the instructions of court special C B I of Lucknow at present this alleged extremist of name.
خفیہ ایجنسیوں کا کہنا ہے کہ ان انتہا پسندوں کو پاکستان اور بنگلہ دیش میں ’دہشت گردانہ‘ حملوں کی ٹریننگ دی گئی ہے۔ ریاست کےڈائرکٹر جنرل پولیس وکرم سنگھ نے بتایا کہ ان افراد نے کم سے کم ایک سو پچاس نوجوانوں کوسرحد پار لے جاکر دہشت گردی کی ٹریننگ دلائی ہے۔	Secret service says that to these extremist is given training of 'terror [grdanh] `attacks in Pakistan and Bangladesh. General police Vikram Singh director of state told that these individuals have at least taken has assigned training of terrorism after across the border gone to one hundred fifty young men.	Secret service says that to these extremist is given training of 'terror گردانـــہ `attacks in Pakistan and Bangladesh. General police Vikram Singh director of state told that these individuals have at least taken has assigned training of terrorism after across the border gone to one hundred fifty young men.

Figure 4 Parallel Text in Spreadsheet Format