ADAPTATION OF THE DARPA MACHINE TRANSLATION EVALUATION PARADIGM TO END-TO-END SYSTEMS

John S. White
Theresa A. O'Connell

Research and Development
Litton PRC
1500 PRC Drive
McLean, VA 22102-5050
tel: 703-556-1899

Abstract

The Defense Advanced Research Projects Agency (DARPA) Machine Translation (MT) Initiative spanned four years. One outcome of this effort was a methodology for evaluating the core technology of MT systems which differ widely in approach, maturity, platform, and language combination. This black box methodology, which proved capable of measuring performance of such diverse systems, used methods which collected human judgments on MT output; expressed these judgments in a quantitative form; and assessed them to evaluate the quality of MT output. This paper investigates the feasibility of applying this methodology to MT efforts beyond the original DARPA goals, into evaluation of MT as part of end-to-end multilingual document understanding systems.

0. Introduction

The Advanced Research Projects Agency (DARPA) Machine Translation (MT) Initiative was a four-year, comprehensive effort intended to develop new MT approaches and translation algorithms. In addition, the initiative developed methodologies for evaluating MT systems, which were appropriate for measuring the progress of core translation algorithms (White et al., 1994). Today, practical application of MT in the workplace introduces new challenges to MT evaluation. Work currently underway at the Federal Intelligent Document Understanding (IDU) Laboratory has identified the need for innovative approaches to evaluate MT systems which function as components of larger end-to-end systems (Craig, 1995). This paper describes some of the challenges this effort faces. It offers modifications to the DARPA MT Evaluation Methodology which speak to the wider needs of applied MT.

The ultimate functionality of MT will hinge on its ability to operate within systems performing multi-step text handling operations, such as optical character recognition (OCR), detection, and extraction. The role of MT in these systems will be quite different from the conventional view of MT as a standalone system. The burdens added to the MT system include: interoperability with other components of the end-to-end system; quality of the output of components which

feed into the MT component; impact of the quality of MT output on other system components; and human factors issues which bear upon human computer interaction. The principles of the DARPA MT Evaluation methodology may be applied to MT components in this context, and appear, with some modification, to be germane to several of the end-to-end text handling tasks.

1. The DARPA Methodology

A significant goal of the DARPA MT Evaluation program was to develop a set of evaluation processes for a general standard. The methodology developed is based on human judgments, taking advantage of the highly subjective nature of opinions about the accuracy of all translation (including human expert). The DARPA solution to handling the subjectivity of human assessments was to decompose these judgments into a large sample of small units, focused and controlled by separate evaluations for adequacy, informativeness, and fluency (White et al., 1995). Expert human translations served as ground truth. This methodology successfully assessed the quality of the output of MT systems as stand-alone entities.

1.1. Adequacy Evaluation

Adequacy is the degree to which a machine translation contains the meaning expressed in an expert human translation of the same source passage. Expert reference translations are divided along syntactic constituent lines into meaningful sentence fragments. The evaluator examines each fragment of the reference translation in turn, then searches for the equivalent meaning in individual paragraphs of the translation under evaluation. Judgments are rendered on a five to one degrading scale where five indicates that all of the meaning is present and one indicates that little or none of the meaning is present.

1.2. Informativeness Evaluation

Informativeness is the degree to which a machine translation provides needed information to its user. The information need is expressed in the form of multiple choice questions. Evaluators read a translated passage and select the best multiple choice answers to a series of questions about the content of the passage. Evaluators are instructed not to apply real world knowledge, to base their answers solely on information stated in the text under evaluation. Although the format of this evaluation resembles a reading comprehension test, it is the ability of the text to inform, not the evaluator's ability to read, that is assessed (Church et al., 1991).

Six multiple choice questions are developed for each text. These are based on the same expert human translations that serve as authority versions in the adequacy evaluation. Questions are written in such a way that no inferencing is required on the part of the evaluators. For each question, the evaluator has six choices: four multiple choice answers; "none of the above"; and "cannot be determined." "None of the above" denotes that the answer is clearly present in the text, but is not one of the choices offered. "Cannot be determined" covers those cases where the evaluator cannot tell, based on reading the text, what the answer is.

1.3. Fluency Evaluation

The fluency measure combines two criteria: the well-formedness of a translation and the degree to which each sentence makes sense in context. Evaluators render intuitive judgments on a sentence by sentence basis. These are expressed on a degrading scale of five to one. Evaluators are instructed to trust their early reaction to each sentence, to avoid analyzing the structure of the sentences. This evaluation differs from the others in that it does not depend on ground truth expert translations.

2. The Federal IDU Lab Challenge

Testing and evaluation at the Federal IDU Lab focus on the usability of end-to-end systems. Software applications in any of the document understanding technologies must demonstrate the potential of usefulness to analysts and other information consumers. Thus the Lab charter extends well beyond testing the performance and functionality of such software against current standards, vendor claims, and/or government contract requirements. The fundamental role of the IDU Lab is to perform testing for the usability of these applications in a realistic, end-to-end functional setting, in which individual IDU applications are integrated together and with other automation tools in the user environment.

2.1. A Typical IDU Scenario.

An end-to-end Intelligent Document Understanding (IDU) system will convert hardcopy in a source language into machine-readable documents in a target language and make relevant materials available to users for information detection and extraction. Such a system may be fully integrated, partially integrated or simply offer the user the opportunity to exit one component and open its output in the next component. The unifying element is the passing of output from one component to the next for processing.

MT may interact with OCR, extraction, detection, and other components in a variety of ways, depending on the intended function(s) of the system. The scenario offered in Figure 1 reflects a typical interaction of functions. This sequence directs the passage of foreign language hardcopy through six steps. Its final goal is extraction of needed information in English. The ultimate beneficiary of this IDU processing is a monolingual information analyst. This analyst has a need for up-to-the minute information.

2.1.1. Step One: OCR

At the front end of this IDU system is an OCR component capable of handling foreign language character sets. Hardcopies of documents which have some probability of meeting analysts' information needs undergo scanning and OCR. A monolingual intermediary performs this task. However, a low accuracy rate in the OCR output will lower the quality of FAMT. Post-editing of OCR output by language experts may be necessary for character sets such as Chinese where recognition quality is poor or in those cases where the source document was degraded by wear, poor paper quality, speckling, etc.

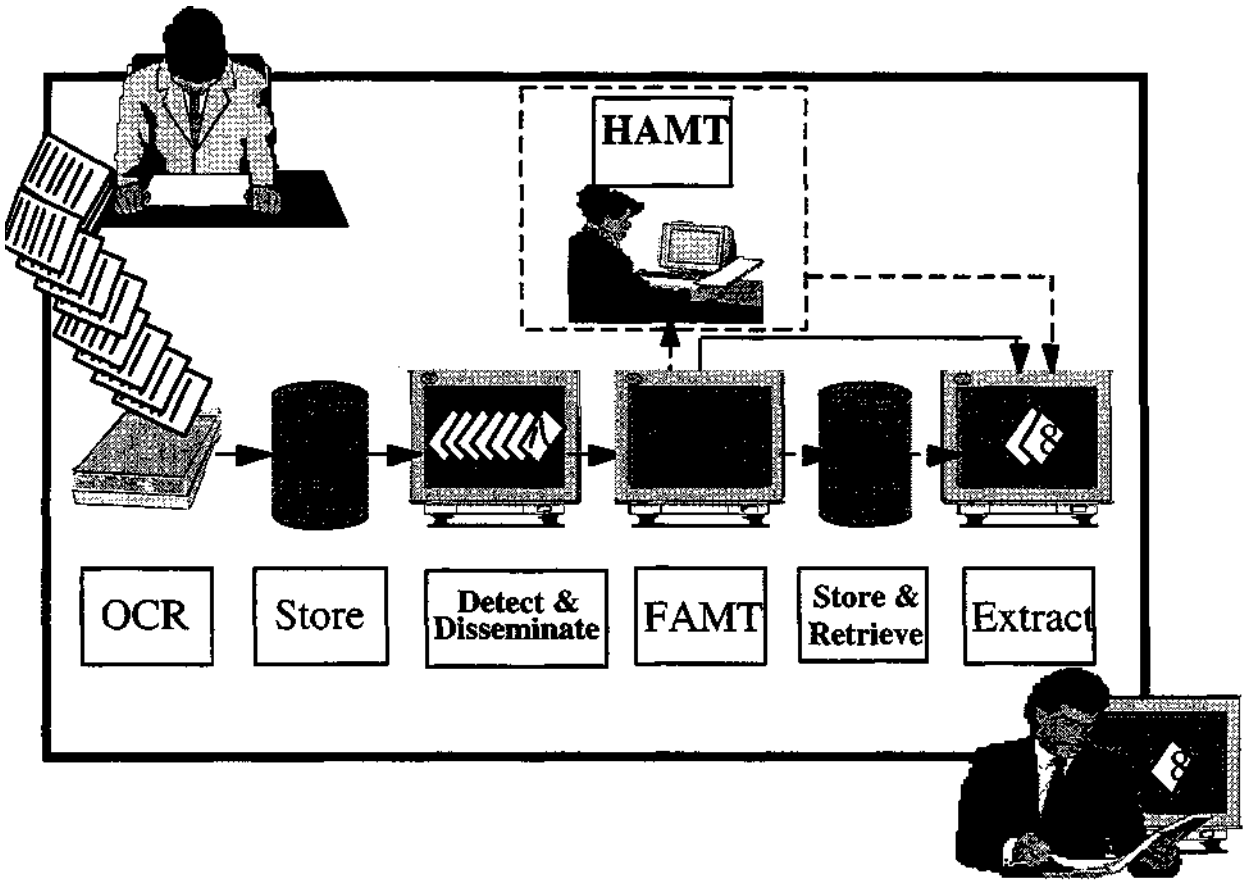

Figure 1. Model of an end-to-end IDU system.

2.1.2. Step Two: Storage

The machine-readable outputs of the OCR are automatically tagged and stored in the source language. Tagging identifies important elements within the document such as its source, date of acquisition, title, and body to facilitate retrieval. Identification of key areas and mark-up of the source text can be automatically generated and therefore administered by a monolingual intermediary.

2.1.3. Step Three: Detection and Dissemination

A monolingual information analyst executes need-based detection, capable of multilingual matching. This process identifies and retrieves documents of probable relevance to the individual analyst (full MT is often proposed for this function as well). A variation on this step automatically disseminates retrieval to the analyst. This detection is based upon stored profiles of users' needs. A trained professional is required to compose and maintain these profiles because the quality of the detection depends upon the quality of the profiles. Detection may

draw upon clues provided by the tagging or it may depend upon full-text retrieval based on strategic words within the document.

2.1.4. Step Four: FAMT

The analyst runs disseminated documents through fully automatic MT (FAMT) into English. The analyst quickly examines the output for relevance and to determine if a language expert should use MT as a tool to produce a more polished translation. Another possible goal of the FAMT is to provide some point of needed information to the analyst. The qualitative requirements of the MT increase according to the analyst's needs: quick gisting (adequacy); polished final product (fluency); or isolated point(s) of information (informativeness). FAMT is used here because the monolingual analyst does not have the expertise to interact with an MT system and refine the translation. If fluency is the goal, the services of a language expert are required. This expert will benefit from a human-assisted MT (HAMT) system as a tool in producing a polished translation.

2.1.5. Step Five: Storage and Detection

Translations may also be stored for later detection by other analysts. There is a correlation between sophistication (adequacy, fluency and informativeness) of the MT and the required sophistication of the detection engine. For example, if the analyst, searching for a specific topic or for a specific point of information, executes a simple keyword search and the one keyword specified is present, detection will be successful. Conversely, if this one word is not present, detection fails. In a document which is both fluent and adequate, more sophisticated search engines which execute meaning-based retrieval can be used.

2.1.6. Step Six: Extraction

Finally, the analyst initiates automatic extraction of needed information, resulting in filled templates which contain, in English, the essential elements of information from sources in many languages. Extraction may also be performed after human intervention with the MT.

2.2. Evaluation Issues

In a functional design like the one described above, MT interacts at one or more points with other software systems which render natural language information into other forms. Each component, OCR, MT, extraction, detection, has the potential of affecting data in ways that may degrade the overall information content, even while enhancing its overall usefulness to a monolingual user. In this context, MT evaluation must focus on the contribution of the MT component to the overall system output, while filtering out degrading effects from other language handling components. The evaluation must be able to distinguish between phenomena that the MT system would have handled correctly in other configurations (or alone), from phenomena that the MT system cannot handle in any case. The goal of MT evaluation in this context must include the ability to gauge the performance of the MT system itself, in order to determine whether the best improvement path for the overall system includes improvement or replacement of the MT component.

Finally, the assessment of performance contribution of MT (or any other component) is moot if the end-to-end system is unusable in the intended setting and by the intended users.

2.2.1. Impact of MT Output on Other Components

It may not be possible, or desirable, to fully evaluate every component of the system for the sole purpose of determining the effectiveness of the translation component. There are, however, predictable interactions among language understanding components, whose characteristics can predict some types of errors (with certain MT designs).

MT —> Detection: In the scenario described above, MT can possibly be used as an input to detection, i.e., by translating all documents in the stream and then feeding those outputs into a common, monolingual profile detection system. Poor MT on detection will obviously impair the potential for high precision, high recall detection. Presumably, this failing is somewhat recoverable, since the MT may not fail on all occurrences of the search string or concept. For example, MT which is run on the output of OCR may be unable to handle a string of characters in one case because of OCR degradation, but succeed with the same string correctly recognized in another case. However, the overhead of translating every document prior to any relevance culling is rather high.

MT —> Extraction: Poor translation affects extraction more severely than detection. Extraction requires a strong lexical, syntactic, and/or semantic fit to template fills. Further, extraction excesses typically involve merging of references (so that actions of Al Gore and "the Vice President" can be attributed to the same individual). Poor MT will seriously disrupt the anaphoric cues that make merging possible.

2.2.2. Impact of Input Quality on MT

It has been hypothesized that some language understanding systems (extraction and detection, for example) may survive poor OCR output, the reasoning being that enough information will be accurately recognized to permit an acceptable level of recall and precision. MT, however, is much more fragile with respect to ill-formed input, particularly if the MT analysis algorithm centers around parsing of natural language input at a sentential level. MT errors measured in evaluation may actually be within the coverage of the MT design, but negatively influenced by OCR errors.

3. Adaptation of MT Methodology

The above discussion raises the obvious differences between the standalone research evaluations and the end-to-end evaluation of MT in-situ. The participants in such evaluations are also different. More than proficiency in English is required; these evaluations will involve functional experts who are stakeholders in the automation of text handling processes. Nevertheless, the similarities are significant: the evaluation of MT in-situ is still a black-box test (though it constitutes a glass box evaluation of the overall IDU system); further, the participants are still monolinguals. The Federal IDU Lab approach to evaluation takes advantage of these similarities

in adapting the original DARPA MT evaluation methodologies. To segregate the impact of possible degradation during OCR in the scenario described above, two versions of the input will be assessed for each measure. The first will undergo optical character recognition prior to detection and MT and the second version will not.

The goal of the DARPA MT Initiative was to promote radical advances in the core technology that transforms a string in one language into a string in another language, hence, the focus on the core technology and the solution of black box testing. However, in those workplaces where MT is an important tool used to achieve defined tasks with explicit goals, end-usability issues arise. Human factors issues which were inappropriate to the DARPA black box methodology come into play and the need arises to incorporate usability evaluation into the assessment methodology.

3.1. Adaptation of Adequacy

The adequacy measure can apply to any document understanding system that includes a detection and/or browsing requirement. Like those processes, adequacy consists of making judgments based on the presence of information in processed documents. The adequacy measure, essentially unchanged from the DARPA version, will be performed directly on MT output, i.e., before it has been further processed by a downstream application.

3.2. Adaptation of Informativeness

As adequacy is relevant to detection, informativeness most resembles extraction. In extraction exercises, like those of the Message Understanding Conference (MUC) series (Sundheim, 1995), text is automatically searched for pertinent information to fill templates which specify particular information (date of report, type of incident, etc.). Rather than modeling users' needs by writing questions, for the end-to-end system it appears that the comparable measure of a translation's ability to provide needed information can be determined at extraction time. Information needs previously represented by questions become slots in the extraction frame. In the case of the end-to-end example above, this output appears at the end of the entire process. Here, extractions performed on both English source texts and translations into English are compared against ground truth template fills; the extent to which the machine translations score differently from the English source texts will be the measure of the MT system's contribution to the informativeness of the whole system.

3.3. Adaptation of Fluency

The measure of fluency depends entirely on the use of MT in the end-to-end system, and is thus most driven by users' needs. There are some information browsing and detection scenarios which, unlike the current example, use MT to create multilingual keyword search queries. In those scenarios, fluency may not be as crucial as in the current one, where MT is a later process. Here, MT must at least have sufficient fluency for verifying that items are of interest. Moreover, since fluency is often a black-box indicator of the successful functioning of MT processes such as parsing, it is worth evaluating: a (linguistic) MT output whose parse has failed will not represent semantic relationships correctly, which will in turn have negative impacts upon extraction in the

scenario above. Control (expert translations) and MT output are evaluated for fluency as in the DARPA methodology, then the same set is run through the detection component of the system. The results should correlate, i.e., texts with low scores should be detected less accurately. The measure can also be applied to documents sent through the end-to-end process and compared to the first evaluation, to factor out effects of the OCR.

3.4 Usability of the MT Component

Thus far, this discussion of system evaluation has focused on the inputs and outputs of components, principally on the adequacy, informativeness and fluency of MT. MT has been presented as part of a larger system and the notion of an end-user has been added to the DARPA MT Evaluation paradigm. With the introduction of the user, there arises a need to evaluate the usability of system components which require human interaction.

Determination of usability starts with an analysis of users' needs and goals. In the case of the MT component, these do more than assign a burden of high quality output. For FAMT, they introduce the requirement that the user be able to successfully initiate MT to achieve the desired output. The user in the scenario described above is a monolingual analyst who must quickly and easily pass a foreign language text through the MT system to facilitate access to needed information. There is also a second class of user with more complex requirements of the system. This is the language expert who will use MT as a tool to develop a fully polished translation capable of undergoing sophisticated and sensitive extraction procedures.

Usability evaluation examines utility, user satisfaction, and the return on user's time invested by assessing how well end-user interface (EUI) features and functions support users' goals and needs. At the simplest level, the user needs to be able to quickly and accurately introduce the output of the prior component, in this case OCR, into the MT component and initiate its FAMT. For example, in a graphical user interface (GUI), a logical menu structure or buttons with labels which clearly set out their function support speedy and error-free task initiation. The language expert requires an EUI that facilitates editing, for example by providing easy access to appropriate lexicons.

While MT component performance time is certainly a factor, it is segregated from user performance time to better focus on user interaction. Usability measures include the length of time it takes a user to successfully initiate a task and the number of errors which occur during task initiation. User satisfaction is surveyed by a questionnaire with quantitative answers converted to a net satisfaction index for each interface feature or function evaluated. The degree to which passage through the MT component has impacted access to needed information is an assessment of the system as well as the MT component. The value of new benefits is ultimately assessed by determining the worth of the filled frames which result from extraction: has the analyst received the appropriate quantity of relevant information and is this information more relevant to his needs?

4. Conclusion

The ultimate use of MT, like the ultimate use of spell-checkers or sorting algorithms, will be as embedded tools within larger, usable, function-oriented systems. Evaluation of the performance of core MT algorithms will continue to be vitally important. At the same time, evaluations of MT systems as components of end-to-end systems must be relevant to their functional context, in terms of performance and usability. The MT component plays a pivotal role in an IDU system; the success or failure of downstream processes depend upon the quality of the MT output. This motivates a need for evaluation of the MT output and of the human-computer interaction which initiates it. The methods described here are intended to serve those ends, and, at the same time, to remain comparable to the standalone MT evaluation measures. Application of the DARPA MT Evaluation Methodology at the Federal IDU Lab will determine its adaptability to a different paradigm, evaluating MT components within end-to-end text processing systems.

References

Church, K., and E. Hovy. 1991. "Good Applications for Crummy Machine Translation." In J.G. Neal and S.M. Walter (eds.) *Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop.* Rome Laboratory Final Technical Report RLTR91362.

Craig, G. 1995. Federal Intelligent Document Understanding (IDU) Laboratory. *Proceedings of AIPA95.* Tyson's Corner, VA., March 1995: Automatic Information Processing Association Steering Group.

Sundheim, B. 1995. *The Proceedings of the Fifth DARPA Message Understanding Evaluation and Conference.* Morgan Kaufmann.

White, J.S., and T.A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas (AMTA).*

White, J.S, T.A. O'Connell, and F. O'Mara. 1995. Evaluation Methodologies in the ARPA Machine Translation Initiative. *Proceedings of AIPA95.* Tyson's Corner, VA., March 1995: Automatic Information Processing Association Steering Group.