Task-based MT Evaluation:

From Who/When/Where Extraction to Event Understanding

Jamal Laoudi+^, Calandra R. Tate*^, Clare R. Voss^

+ARTI, Inc.
Alexandria, VA
jlaoudi@arl.army.mil

*Dept of Mathematics
U. of Maryland, College Park, MD
ctate@arl.army.mil

^Multilingual Computing Group
Army Research Lab, Adelphi, MD
voss@arl.army.mil

Abstract

Task-based machine translation (MT) evaluation asks, how well do people perform text-handling tasks given MT output? This method of evaluation yields an *extrinsic* assessment of an MT engine, in terms of users' task performance on MT output. While this method is time-consuming, its key advantage is that MT users and stakeholders understand how to interpret the assessment results. Prior experiments showed that subjects can extract individual who-, when-, and where-type elements of information from MT output passages that were not especially fluent. This paper presents the results of a pilot study to assess a slightly more complex task: when given such wh-items *already identified* in an MT output passage, how well can subjects properly select from and place these items into wh-typed slots to complete a sentence-template about the passage's event? The results of the pilot with nearly sixty subjects, while only preliminary, indicate that this task was extremely challenging: given six test templates to complete, half of the subjects had *no* completely correct templates and 42% had exactly one completely correct template. The provisional interpretation of this pilot study is that event-based template completion defines a task ceiling, against which to evaluate future improvements on MT engines.

1. Introduction

Task-based machine translation (MT) evaluation asks, how well do people perform text-handling tasks given MT output? This method of evaluation yields an *extrinsic* assessment of an MT engine, in terms of the user's task performance on MT output. While this method is time-consuming, its key advantage is that MT users and stakeholders understand how to interpret the assessment results.[1]

In this paper, we present the results of an online pilot study evaluating how effectively three types of Arabic-English MT engines[2] support MT users in understanding events in newswire passages. Our goal was to identify a task near the ceiling of MT support, against which to calibrate future MT engine improvements. We knew from annotators in another evaluation study that, with practice, they could extract individual who-, when-, and where-type elements of information from MT output passages that were not especially fluent. The pilot study was to assess whether subjects, when given these items pre-identified in MT output, could piece them together to complete a sentence about the passage's event, the "who did what to whom when and where."

Nearly sixty subjects were given MT output with wh-units of content *already* identified. The results are only preliminary and indicate that this task was extremely challenging: given six test templates to complete, half of the subjects had *no* completely correct templates and 42% had exactly one completely correct template. The best performance came from 8% of the subjects who had only two completely correct templates. The provisional interpretation of this pilot is that event-based template completion define a task ceiling, against which to evaluate future improvements on MT engines.

2. Background

Extrinsic, task-based evaluation of MT engines has long been of interest to those who seek automated support tools to expedite their decision-making tasks (Spaerck-Jones and Gallier, 1996). In the late 1990's two new MT research trends emerged, furthering interest in extrinsic metrics: task-based experiments were being conducted by MT developers on their own engines (Resnik, 1997; Levin et al., 1999), and task-based experiments assuming an ordering of task difficulty were being proposed by users on text-handling tasks (Taylor & White 1998).

Task	Description of Task
Publishing	Produce technically correct document in fluent English
Gisting	Produce a summary of the document
Extraction	For documents of interest, capture specified key Information
Triage	For documents determined to be of interest, rank by importance
Detection	Find documents of interest
Filtering	Discard irrelevant documents

Table 1: Proposed Hierarchy of Text-Handling Tasks (Taylor & White, 1998)

Then, with the introduction of several automatic MT metrics[3] demonstrating both the vitality of MT evaluation as a research area of its own and the impact of metrics on the MT development cycle, MT stakeholders began

[1] By contrast, with the automated metrics where ngram string-based algorithms match MT output text against one or more human reference translations, MT users are unclear what the scores mean (Callison-Burch, Osborne, & Koehn, 2006) and how the scores are related to their tasks (Tate, 2005).

[2] MT-1 was rule-based, MT-2 statistically-trained, and MT-3 substitution-based with lexicon and morphological analyzer.

[3] Such as BLEU (Papineni et al. 2002), GTM (Melamed, Green, & Turian, 2003), METEOR (Lavie, Sagae, & Jayaraman, 2004), and TER (Snover et al. 2005).

funding research experiments in task-based assessment of MT engines, to address users' needs.[4]

2.1. Selection of Tasks

After reviewing Taylor & White's hierarchy of tasks, shown in Table 1, and examining the MT output of several engines, we designed three experiments to test for: (i) one task as a lower-bound for a *shared capability*, that all the selected different types of Arabic-English MT engines could support, (ii) one task as an intermediate challenge, that one or two engines would support but another one would likely not, and (iii) one task as an upper-bound for a *shared limitation*, that none of the selected engines could yet support.

A small, prior pilot experiment to evaluate Arabic-English MT engines for document-exploitation tasks indicated that subjects could extract some named entities and event participants from noisy MT output, but they could not readily identify relations within events (Voss, 2002). This led to the selection, for task (ii), of wh-item extraction, a task between event-level analysis and named-entity recognition (see Table 2). This report focuses on the pilot experiment conducted for task (iii).[5]

Levels of Extraction	**Description**
Deep	***Event identification*** (scenarios): the ability to identify an incident type and report all pertinent information
task (iii)	***Event completion:*** identify argument and adjunct relations among who-, when-, where-type elements of information
Intermediate	***Relationship identification*** (e.g., member-of, associate-of, phone-number-for)
task (ii)	***Wh-item extraction***: Identification of *who-types* (people, roles, organizations, companies, groups of people, government), *when-types* (dates, times, duration or frequency in time, proper names for days & common nouns referring to time periods), *where-types* (geographic regions, facilities, buildings, landmarks, spatial relations, distances, paths)
Shallow	***Named entity recognition:*** isolation of names of people, places, organizations, dates, locations

Table 2: Multiple Levels of Extraction (Taylor & White, 1998), with extra rows inserted for Event completion, task (iii), and for Wh-item extraction, task (ii).

The primary objective of task (iii) was to develop specifications for and conduct a pilot test on a text-handling operation where subjects would identify "higher-order" relations among phrases at an event level in MT output texts, i.e., relations that were linguistically more complex than those extracted in task (ii), without inferring information not explicitly in the text.

In practical terms, the challenge---given the varying levels of accuracy and fluency in the output of the project's three MT engines---was to ensure that the text-handling operation in task (iii) met the "higher-order" linguistic complexity requirement without being so difficult that subjects would become discouraged and give up when reading the MT outputs. Previous experience had taught us that subjects lose focus on a task and answer randomly when they believe that they cannot perform the operations that they have been trained to do, thereby confounding any assessment of task accuracy (how well can the task be done?) with subject motivation (will the subjects do the task?).

2.2. Selection of MT Systems

In conjunction with the project sponsor, three distinct types of MT engines were selected as representative of three development models, varying in required funding, time, and linguistic resources:

- MT-1, a rule-based engine with handcrafted lexicons and symbolic linguistic processing components (e.g., morphological analyzer, parser)
- MT-2, a statistical engine trained on large quantities of monolingual and parallel Arabic-English texts, but with no traditional, symbolic linguistic processing components
- MT-3, a substitution-based engine that relies entirely on a pattern-matching algorithm with a lexicon and morphological analyzer to translate matched strings into English phrases, replacing the former with the latter, leaving the original Arabic word order unchanged except as occurs locally within the substituted phrases.

3. Pilot Experiment

3.1. Task Description

For the pilot, subjects were first trained on the task with English-original texts and then with MT-output texts, to become familiar both with the software and with the irregularities of MT output. While the subjects in the pilot had performed extraction-like tasks prior to this study, many had had no previous experience working with MT or MT output. Following the training phone, subjects immediately went on to the evaluation phase.

During their training, subjects had the opportunity to practice the task and receive feedback on their responses. Figure 1 shows a sample screen from the pilot. The textbox at the top of the screen holds the MT output text with the wh-type phrases already highlighted. Beneath the textbox is the template with wh-typed slots to be filled. Subjects are instructed (1) to read the document, as displayed in the topmost part of the figure, where the wh-items are color-coded consistently with Who-items yellow, Where-items blue, and When-items purple, (2) to read the accompanying template, positioned below the textbox, where the open slots are color-coded and "typed" with the name of their wh-type, and (3) to decide which Who-, When-, and Where-items marked in the document belong in which slots in the given template.

They need to learn that they can only select one wh-item at a time from the colored (marked) items by clicking on it in the text --- clicking on uncolored text is ignored by

[4] For example, see the 2005 broad agency announcement (BAA) for the Global Autonomous Language Exploitation program (GALE) released by DARPA, a US government funding agency.

[5] Prior research for task (i) was a detection-level, topic categorization pilot by Tate, Lee, Voss (2003). Details of tasks (i)-(iii) are in Voss et al (2006).

the software--- and then clicking on one similarly color-coded slot in the template, to copy that item's text into the slot. They must fill each of the labeled slots in the template with exactly one wh-item of the type specified by the slot. (To change the content of the slot, the subject only needs to click the preferred wh-item in the main texbox and then click on the slot to-be-changed in the template: this automatically replaces the slot content with the just-selected wh-item, as long as it is correctly typed. Subjects then continue clicking to copy phrases into the other slots in the template until all slots were filled. All the slots in a template must be filled before pressing the 'Next' button to move on to the next document.

To facilitate the subjects' tracking of which document items they had already selected and copied into template slots, the software automatically bolds the text of the wh-item selected in the document immediately after the subjects completed the copying.

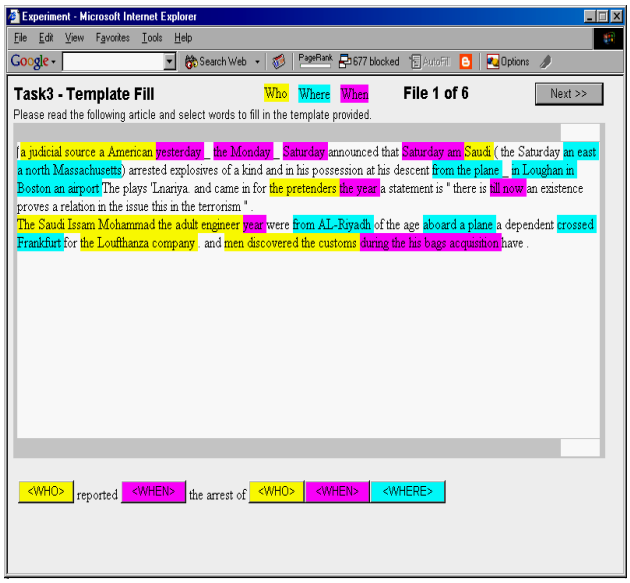

Figure 1. Sample Screen: MT output in textbox, template with slots below.

[*An American judicial source*] reported [*yesterday, Monday,*] the arrest of [*The Saudi man, Issan Mohamed Al-Mohandiss*] [*Saturday*] [*at Logan airport*] [*in Boston Massachusetts*]

Figure 2. Template with Reference Translation answers

While the task is relatively straightforward with coherent text, such as the English-original presented in training, it proved quite challenging with machine-translated text. For example, Figure 2 shows that the subject who-item for the template in Figure 1 corresponds to the phase "An American judicial source", while the correct, corresponding phrase available in the MT output is "a judicial source a American". Thus the training phase also included examples of MT output so that subjects could see that acceptable wh-items were typically NOT grammatical or fluent.

The evaluation phase followed immediately after the training phase. Subjects remained at their desktop PCs and used the same software and procedures as in the training phase, except that they were given no verbal instructions, they were not allowed to ask experiment administrators for assistance, and they received no online feedback to their responses.

3.2. Experiment Design and Data

For this task, all source language documents for the document collection were selected in a first pass in conjunction with reading their reference translation and creating an event template for each with who-, where-, when-type slots. After each document was translated once by each of three MT systems, the MT output texts were examined to be sure that the wh-items for completing the template slots came through in the translation.

The following two constraints on the experimental design ensured a balance in the number of MT engine and document viewings for each subject:

Constraint 1: No subject should view the same document more than once, i.e., the same source document as translated by more than one MT system.

Constraint 2: Each subject should view an equal number of documents translated by each MT system, at least two.

The pilot involved three MT systems (MT) and six source language documents (DocId), each with their own template, yielding a 18-document collection indexed by all MT x DocID combinations. The total number of ***Cases*** for recording results, i.e., instances of translated documents viewed by subjects who completed the templates, was equal to the product of the number of docID's, the number of MT systems, and the number of subjects who viewed each translated document.

Subject	X	Y	Z
	MT-2 Doc1	MT-1 Doc 1	MT-3 Doc 1
	MT-2 Doc 2	MT-1 Doc 2	MT-3 Doc 2
	MT-3 Doc 3	MT-2 Doc 3	MT-1 Doc 3
	MT-3 Doc 4	MT-2 Doc 4	MT-1 Doc 4
	MT-1 Doc 5	MT-3 Doc 5	MT-2 Doc 5
	MT-1 Doc 6	MT-3 Doc 6	MT-2 Doc 6

Table 3. Three-block design for task (iii), to be randomized as superblock for 3 viewing sequences

The experimental design was based on the three-block document set in Table 3. Each translated document was assigned a unique identifier, combining the translation engine id (MT-1, MT-2, or MT-3) and one integer from 1 to 6 for the DocID. The resulting 18 identifiers were distributed across the three blocks (or subject-viewing sequences) in columns X, Y, and Z, with two translated documents for each of the three MT systems in each column.

With the above constraints and a final count of 59 subjects each with 6 viewings, there were a total of 354 viewings in this experiment. This included 118 total viewings of each MT system, 59 total viewings (across translations) of each document, and 19 or 20 total viewings of each translated document. The variation in the 19 or 20 viewings occurred because the subject pool was one person short of the full experimental design for 60 subjects, based on 20 superblocks for three subject viewing sequences each, as originally anticipated.

4. Analyses and Results

The subjects' responses were scored against a set of answer-templates created from the reference translations. For each of the 354 ***cases*** of data collected in this pilot**,** corresponding to one translated document viewed by one subject and accompanying template completed by that subject, we created three sets of scores. For each completed template,

- for *all slots,* we recorded whether the template was fully correct, i.e., all of its slots were correct, or not.
- for *only who-type slots*, we recorded whether ALL such slots in the template were correct (scored A), or one or more were not (scored non-A)
- for *only who-type slots,* we recorded whether NONE of the slots in the template were correct (scored N), or one or more were correct (scored non-N).

4.1. Fully Correct Templates

We first analyzed subject responses by counting their completely correct templates, i.e., where the selection and placement of all phrases in the slots were correct. The templates in the pilot varied in the number of who-, when- and where-type slots (see Table 4).[6]

As shown in Table 5, out of the 354 total viewings (where 59 subjects each saw 6 translated documents), only 34 viewings produced fully correct templates. Out of the 118 viewings of documents translated by one MT engine, even in the best case, the subjects completed no more than 20% of the templates correctly, and most of these were attributable to a single translated document, doc 2.

Clearly, this task, as defined, developed, and administered for this pilot, proved exceedingly difficult. It is interesting to note that the worst scores were on templates with more than two who-arguments.

	doc 5	**doc 3**	**doc 1**	**doc 4**	**doc 6**	**doc 2**
who	4	3	2	2	2	2
where	0	1	2	1	1	1
when	0	1	0	1	2	0

Table 4. Number of slots to be filled by wh-type in each document-template

	doc 5	**doc 3**	**doc 1**	**doc 4**	**doc 6**	**doc 2**	**Total**
MT2	0	1	0	3	5	14	23/118
MT1	0	0	2	0	4	2	8/118
MT3	0	0	0	0	3	0	3/118
Tot	0/59	1/59	2/59	3/59	12/59	16/59	34/354

Table 5: Number of fully correct template responses

4.2. Who-Slot Analyses

Given how poorly the subjects did on the full templates, we asked next, how well did subjects do on just the Who-type slots? In one analysis, we divided the subjects' template responses for the Who slots into A and non-A categories: the A-level responses had all of the who-items in the subject's template correctly identified and in correct slots, and all other responses were categorized as non-As.

The results for the number of A-Level vs. Non-A Level templates are shown in Table 6, with the MT system variable separated into MT-2 vs. MT-1 and MT-3. Analysis of this table yields a chi-square value of 14.93. Under 1 degree of freedom, this statistic is certainly extreme. Thus, the difference in A-level and non-A level responses between MT-2 and the other two systems is statistically significant.

	A Level	Non-A Level
MT2	48	70
MT1 & MT3	50	186

Table 6. A-Level vs. Non-A Level accuracy in responses

In the third analysis, the subjects' template responses were divided into N and non-N categories: the N-level responses had NONE of the who-items in a subject's template correct, and all other responses were categorized as non-Ns. The results for the number of N-level vs. Non-N level answers are shown in Table 7, with the MT system variable separated into MT-3 vs. MT-1 and MT-2. Analysis of this table yields a chi-square value of 18.47 under 1 degree of freedom. This statistic is again extreme and the difference in N-level and non-N level responses between MT-3 and the other two systems is statistically significant.

	N Level	Non-N Level
MT1 & MT2	13	223
MT3	24	94

Table 7. N-Level vs. Non-N Level accuracy in responses

Thus, the main conclusion of this pilot study is that the experimental procedures developed, tested, and reported on here, yielded the following result: no MT engine appears yet to be adequate to support subjects on task (iii). The subject responses on who-type items suggest that one ranking of the systems, that could be considered along with the others from task (ii), in the order of performance from strongest to weakest is: MT-2, MT-1, and MT-3.

5. Conclusion and Future Work

The pilot was conducted in Feb. 2004 with MT engines developed up until Oct 2003. Subjects were given MT output with wh-units of content *already* identified. The results are only preliminary and indicate that this task was extremely challenging: given six test templates to complete, half of the subjects had *no* completely correct templates and 42% had exactly one completely correct template. The provisional interpretation of this pilot is that event-based template completion defined a task ceiling, against which to evaluate future improvements on MT engines.

We are now examining the outputs of current, upgraded versions of the same Arabic-English MT engines on the same original Arabic texts. We see some

[6] Varying the number and type of slots in the pilot was intentional. It provided some hints, to explore at a later time, as to which combinations are easier and which harder to complete..

limited, within-phrase improvements in fluency and word choice for all three engines, so the overall event-level information remains difficult for MT users to discern.

Table 8 shows just a sample from the beginning of one sentence from one document, as output by MT1 and MT3 in their 2003 and 2005 versions. The cells with * indicate that this wh-item was not found in the translation. The cell with /\ indicates that its when-item is in the cell above, namely that its when-item "yesterday" was located within the who-item. The cell with \/ indicates that the where-item "field" was located within the who-item.

This snapshot of task (ii)-like wh-item highlighting, as limited as it is however, suggests that, for rule-based MT1 and substitution-based MT3, there is room for improvement.

Reference Translation	**MT1 (2003)**	**MT1 (2005)**	**MT3 (2003)**	**MT3 (2005)**
<where> On the field	Two fields	Two fields	\/	On the Ground
<who> an American Officer	an officer a American	a ***yesterday*** American officer	us ***field*** officer	US Officer
<when> yesterday	*	/\	yesterday	*

Table 8. Output of MT1 and MT3 from 2003 and 2005

The Appendix also provides a close, but still quite limited, look at the original and current, upgraded MT2 translations of the document for which subjects were most likely to correctly complete the template. While there are fewer unknown words transliterated and fewer spurious words not clearly lexically related to any words in the source text in the current versions of the engine, we see some phrase-internal word ordering improved, even though the Arabic Verb-Subject-Object syntax is still being carried over into English.

Acknowledgements

The pilot study was part of a larger project funded by the Center for Advanced Study of Language (CASL), University of Maryland, College Park, and the U.S. Army Research Laboratory, Adelphi, Maryland. Software development and experiment administration were provided by ArtisTech, Inc. Sooyon Lee contributed to the computer programming and data analyses.

References

Callison-Burch, C., M. Osborne, P. Koehn (2006). Re-evaluating the Role of BLEU in Machine Translation Research.In *Proceedings of the European Association for Computational Linguistics (EACL).*

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.

Lavie, A., K. Sagae and S. Jayaraman (2004). The Signicance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA'04) . Washington, DC.

Levin, L., B. Bartlog, A. Llitjos, D. Gates, A. Lavie, D. Wallace, T. Watanabe and M. Woszczyna (2000). Lessons Learned from a Task-based Evaluation of Speech-to-speech Machine Translation. Language Resources and Evaluation Conference, Athens, Greece.

Melamed, I. D., R. Green and J. P. Turian (2003). Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL.* Edmonton, Canada.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL*. Philadelphia, PA.

Resnik, P. (1997). Evaluating Multilingual Gisting of Web. In *Proceedings of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*. Stanford, CA.

Snover, M., B.J. Dorr, R.Schwartz, J.Makhoul, L. Micciulla, and R. Weischedel. (2005). A Study of Translation Error Rate with Targeted Human Annotation. LAMP-TR-126. U. of Maryland, College Park.

Spaerck-Jones, K. and J.R. Gallier. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, Berlin.

Tate, C., S. Lee, C. Voss (2003). Task-based MT Evaluation: Tackling Software, Experimental Design, & Statistical Models. In *Proceedings of MT Summit IX Workshop on MT Evaluation.* New Orleans, LA.

Tate, C. (2005). Evaluating Machine Translation Output & Predicting Its Utility. *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* Ann Arbor, MI.

Taylor, K. and J. White (1998). Predicting What MT is Good for: User Judgments and Task Performance. In *Proceedings of Association for Machine Translation in the Americas (AMTA-98),* 364-373, Lansdowne, PA.

Voss, C.R. (2002). MT Evaluation: Measures of Effectiveness in Document Exploitation. Presentation, DARPA TIDES PI Meeting. Santa Monica, CA.

Voss, C., C. Tate, E. Slud, J. Hancock, M. Aguirre, J. Laoudi, S. Lee, S. Shukla, J. Turner, M. Vanni (2006) *TTO 19: Study of MT Utility for Triage.* Technical Report, Army Research Laboratory (ARL)/Center for the Advanced Study of Language (CASL), University of Maryland, College Park, MD. Under review.

Appendix

Here we briefly present the results of running the same Arabic source document from task (iii) through:

- the original MT2 engine (2003) in Figure 3,
- the current (2005), small-memory (500K) version[7] of MT2 in Figure 4, and
- the full, current (2 Gig) version of MT2 in Figure 5.

Reading through each of the machine translations of this passage, curiously enough, we have the impression that the full sequence of events would not be clear to most readers. This suggests even now, a few years after the pilot, that MT engines are not likely to support task (iii) for Arabic to English translation.

A human reference translation of the document is shown in Figure 6 with the accompanying empty template in Figure 7, and a correctly completed template in Figure 8. The reader is encouraged to compare what part of the meaning is preserved and which is lost in the various translations.

A plane landed coming from Paris flying to us safely after a false alarm security, authorities have passengers to conduct further interrogation them after they cause maitf [معطف] woman in a security warning lying. Firefighters passengers to isolated region thirty 184 wmlahw [وملاحو] 12 aircraft from the aircraft then moved to the airport terminal. passenger plane landed a Boeing 767 belonging to Delta in its visit to 43 airport synsynaty [سينسيناتي] North kntaky [كنتاكي] towards late hour.

Figure 3. Original MT2 output (2003)

The original output from MT 2 has two terms introduced into the translation presumably by statistical association: "Firefighters" and the number "thirty". The non-translated words that were transliterated include::

maitf [معطف],
wmlahw [12 وملاحو]
synsynaty [سينسيناتي]
kntaky [كنتاكي]

Sounding out the last two to hear what's there does help the reader figure out what is being written about.

fell a plane coming from Paris oriented to the United States in peace, security warning after a liar, authorities received passengers for further questioning them after the security alert caused middle a woman in a liar. And he brought forth the passengers were drawn to the region remote U?U?U??§?U? plane 184 of the 12 of the plane then transferred to the building of the airport. passenger plane falling back, a Boeing 767 belonging to a type of Delta company journey 43 at the airport ?³U?U??³U?U??§??U? North McDonald late about one hour.

Figure 4. Recent, small-memory MT2 (2005)

In Figure 4, the problematic translation with "Firefighters" and "thirty" (found in original MT-2 output) is gone, while the same strings from the original MT-2 failed to translate. A major issue with this output is the disappearance of the keyword "maitf [معطف]" (coat) that the original MT-2 did not translate, but did provide a transliteration for. Another issue with the output is the creation of the new, possibly spuriously associated translation into the term "McDonald".

Landed a plane coming from Paris heading to the United States peacefully after a warning security a liar, received the authorities passengers for further questioning them after that caused a woman in security alert a liar. And brought out passengers to the region of remote i.e. 184 services aircraft of 12 of the plane then transferred to the building at the airport. And landed passenger aircraft and a Boeing 767 belonging to a company Delta in flight 43 at the airport Cincinnati northern Kentucky late about an hour.

Figure 5. Recent, full-size MT2 (2005)

The recent full-size MT successfully translated more of the source language text and did not introduce any of the previous false associations (as found in Figure 4). However, the reader is not likely to discern the full event of the text passage from this translation: that 184 passengers and 12 crew were the ones detained at the airport because a woman's coat tripped the security alert system. The unknowing reader may also be led to assume that Cincinnati is in Kentucky.

A plane coming to the United States from Paris landed safely after a false alarm, as authorities conducted more interrogations of passengers after a woman's coat set off a false alarm. The 184 passengers and the 12 crew members were taken to a remote location then transported to the airport terminal. The passenger plane Boeing 767 flight 43 landed about an hour late at the Cincinnati airport, north of Kentucky.

Figure 6. Reference Translation for Text

[WHO] relocated [WHO] from the plane [WHERE] and then [WHERE]

Figure 7. Template with Wh-types for Texts in Appendix (Figures 3-6)

[*Authorities*] relocated [*passengers*] from the plane [*to a remote location*] and then [to the airport terminal].

Figure 8. Template with Wh-items from Reference Translation (shown in Figure 6)

[7] The smaller memory version of MT2 allows for a more mobile user to put this engine on a more portable device, such as the FALCon laptop (Fisher and Voss, 1997).