Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic•

Douglas Jones[1], Wade Shen[1], Neil Granoien[2], Martha Herzog[2], Clifford Weinstein[1]

[1]MIT Lincoln Laboratory
244 Wood Street
Lexington, MA, 02420, USA
{DAJ, SWade, CJW}LL.MIT.edu

[2]Foreign Language Center
1759 Lewis Road
Monterey, CA, 93944, USA
{Neil.Granoien, Martha.Herzog}@monterey.army.mil

Keywords: Machine Translation Evaluation, Foreign Language Processing, Usability

Abstract

We present results from an experiment in which educated English-native speakers answered questions from a machine translated version of a standardized Arabic language test. We compare the machine translation (MT) results with professional reference translations as a baseline for the purpose of determining the level of Arabic reading comprehension that current machine translation technology enables an English speaker to achieve. Furthermore, we explore the relationship between the current, broadly accepted automatic measures of performance for machine translation and the Defense Language Proficiency Test, a broadly accepted measure of effectiveness for evaluating foreign language proficiency. In doing so, we intend to help translate MT system performance into terms that are meaningful for satisfying Government foreign language processing requirements. The results of this experiment suggest that machine translation may enable Interagency Language Roundtable Level 2 performance, but is not yet adequate to achieve ILR Level 3. Our results are based on 69 human subjects reading 68 documents and answering 173 questions, giving a total of 4,692 timed document trials and 7,950 question trials. We propose Level 3 as a reasonable near-term target for machine translation research and development.

• This work is sponsored by the Defense Advanced Research Projects Agency and the Defense Language Institute under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

This paper has not already been accepted by and is not currently under review for a journal or another conference, nor will it be submitted for such during IA's review period.

1. Introduction

In 2000, Salim Roukos, Kishore Papineni and colleagues at IBM developed a fast, portable, and easily repeatable measure of MT performance based on N-gram co-occurrence statistics, now known as "BLEU" scores (Bi-Lingual Evaluation Understudy scores; Papineni et al. 2002). BLEU automatically measures the surface-similarity of MT system output to a set of trusted human reference translations for a given document. Building on the work of Roukos and Papineni, George Doddington and Mark Przybocki at NIST adapted and extended the BLEU N-gram scoring technique to form the basis of the official NIST MT evaluation workshops, beginning with a successful dry-run exercise in 2001 and continuing in the MT-02, MT-03 and MT-04 evaluations. Participation in the annual NIST MT evaluations is mandatory for machine translation research sites funded by the DARPA/TIDES program but it has also included many non-DARPA funded sites. Consequently the BLEU/NIST N-gram scoring package has become the common measure of research progress not only for DARPA/TIDES but for MT research world-wide. The past three years have seen an unprecedented advance in MT research results, measured by and in part enabled by the development of these automatic scoring techniques (Wayne 2004).

Our motivation for this pilot study is to translate these technological advances into terms that are meaningful for people who rely on machine translation systems operationally (i.e. the consumers of machine translation). Establishing unbiased measures of machine translation performance is of critical importance to the Government, especially when such measures could influence policy decisions, for instance affecting resources allocated to human foreign language training and foreign language processing technology development. For these measures to be useful, they need to be able to be relatable to real-world tasks and skills.

Our primary complaint about N-gram scores is that they do not relate directly to measures that are easily in-

terpreted by the Government for fulfilling foreign language requirements. Although useful as a measure of progress to the research community, these scores do not allow decision makers to judge the utility of MT for real tasks. To determine the utility of MT technology for such tasks, a different evaluation is needed.

In this study, we adopt a standardized protocol for measuring foreign language proficiency known as the Defense Language Proficiency Test (DLPT) and adapt this protocol for MT evaluation. The DLPT is a high-stakes test used to assess the foreign language proficiency of English speakers. The results of these tests have been used for training and career placement purposes for U.S. military and government personnel for decades. As such, these tests have an accumulated institutional knowledge and history that allows decision makers to relate their results to real-world foreign language capabilities. Since live test materials must be tightly controlled to guarantee the integrity of the tests, we have constructed a new set of DLPT test materials for our experiments.

The DLPT-style questions we use in this study are short-answer questions created by foreign language proficiency testing experts, fluent in Arabic, using original source documents. These questions are designed with strict testing guidelines to test a reader's ability to extract crucial information from the source document. They are highly-specialized and have proven to correlate reliably with other foreign language evaluation techniques, including more labor-intensive direct interviews. Several examples of DLPT questions can be found in the appendix.

1.1 Overview of the ILR Difficulty Levels

The DLPT rates test-takers at different levels of proficiency. These proficiency levels are defined by the Interagency Language Roundtable (ILR) and have been an integral part of foreign language skill assessment in a variety of settings for agencies in the U.S. Government. Each level of proficiency (ranging from 0 to 5) has a corresponding level of text difficulty. Texts of a given difficulty level require that level of proficiency to be understood. A description of an ILR-based text classification scheme can be found in (Child et al., 1993 and Lowe 1999) and on the web (see References); some key points:

- **Level 1 texts:** contain short, discrete, simple sentences; generally pertain to the immediate time frame; often written in an orientational mode; require elementary level reading skill. Example: Newspaper announcements.

- **Level 2 texts:** convey facts with the purpose of exchanging information; do not editorialize on the facts; often written in an instructive mode; require limited working proficiency. Example: Newswire articles.

- **Level 3 texts:** have denser syntax and highly analytic expressions; place greater conceptual demands on the reader; often written in an evaluative mode; may require the reader to 'read between the lines'; require general professional proficiency. Example: newspaper opinion / editorial articles.

- **Level 4 texts**: express creative thinking; assume a relative lack of shared personal information; often involve a highly individualized mode that projects the style of the author; require advanced professional proficiency. Example: essays; political editorials that reformulate social, economic or political policy. (We did not use Level 4 texts in this experiment.)

Test writing experts at DLI assembled a corpus of Arabic texts with passages at levels 1-3 and written DLPT-style questions for use in this experiment. Figure 1 shows translated examples at ILR Levels 1, 2 and 3. Some basic statistics about the corpus are shown in Table 1. A sample Level 2 test item is shown in the Appendix.

Level 1 (Weather Report)

Weather: The weather will be temperate in the northern coasts, getting slightly warmer in Lower Egypt and Cairo, hot in Central Egypt, very hot in Upper Egypt and mild in all other areas during daytime. …

Level 2 (Newspaper Article)

Egyptian security forces at Cairo Airport prevented the retired actress Sabreen from traveling to the Kingdom of Saudi Arabia with her husband Tarek Galal Mahmoud. Sabreen was trying to travel to Saudi Arabia the day before yesterday with her husband when Egyptian authorities discovered that the actress, whose full name is Sabreen Yaseen Mahmoud Abdullah and who resides in Old Cairo's Elmanial quarter, is among those barred from leaving the country….

Level 3 (Political Commentary)

Declarations by the Minister of Foreign Affairs in the "appointed" Iraqi government threatening to allow the American forces that occupy Iraq to launch attacks on neighboring countries seems to be in response to their support of the Iraqi resistance. It is logical for a new regime to lack political and diplomatic experience. However, building bridges with others is a matter of common sense which does not need any special experience. …

Figure 1: MT-DLPT Corpus Sample

1.2 The MT-DLPT

The DLPT for reading proficiency is a highly refined and widely accepted test for determining the language proficiency of foreign language students. Thus the DLPT was adopted as a model for measuring the practical utility of machine translation systems. In order to benefit from the experience and intellectual capital invested in the DLPT, a modification of it was created that measures the adequacy of language translations rather than the proficiency of foreign language students. This was done by replacing the foreign language text with an English translation of it and by replacing the foreign language student with a native speaker of English who does not speak Arabic. Assuming that the English speaker is fluent and of college-level education, then the answers to DLPT questions reflect the adequacy of the translation instead of the foreign language proficiency of the student. This modification of the DLPT we call the "MT-DLPT". The next section describes the experiment that was conducted to measure MT adequacy.

2. Experiment

Using the MT-DLPT, we have conducted an experiment that measures machine translation quality as a function of (1) text passage comprehension, as measured by short-answer questions accuracy and (2) the time taken to complete a test item (a passage + its corresponding questions). To make the test more robust to variations in machine translation quality within and across passages, we constructed a test using 68 text passages. Subjects were allowed a full eight-hour day to complete the experiment.

2.1 Participants

84 participants from MIT and the surrounding community were recruited for participation in the experiments. All were native speakers of English. Participants were excluded if they had significant knowledge of the source language, since they may be more able to understand and compensate for awkward translations of their language.

2.2 Dual Use Arabic Language Materials

A set of 75 Arabic DLPT test items were created at the DLI Foreign language Center as dual use materials, both for use in our machine translation study and for use in foreign language training at DLI. A document pool of approximately 22,000 words, balanced across three difficulty levels (Levels 1, 2, and 3) was selected from a variety of written Arabic media sources. Four sets of high quality English translations were then produced by different translation services. A final translation was created from an adjudication of the four translations. Comprehension questions, in English, were then written for a subset of the document pool in order to balance the number questions at each level and the amount of time a subject would need to complete the entire test. Table 1 illustrates a balance of passages and questions across the difficulty levels. The text passage size generally increases with difficulty.

Table 1 MT-DLPT Corpus Statistics

	Document Pool		MT-DLPT Test Documents		
	Docs	**Words**	**Docs**	**Words**	**Questions**
Level 1	101	5,138	30	1,203	61
Level 2	35	5,953	20	3,485	59
Level 3	18	10,964	18	9,703	53
TOTAL	154	22,055	68	14,391	173

68 passages were used in the main experiment (total across Level 1, 2 and 3). Versions of each passage were created for each of the following experimental conditions: MT (machine translation) and HT (high quality professional reference translation produced by human translators). We used a Latin Square design in which all experimental conditions are balanced across subjects and each subject sees each item in exactly one condition. The order of the texts in the experiment followed the structure of the DLPT, in which texts appear at increasing levels of difficulty. Within each level, texts were pseudo-randomly selected for each subject on an individual basis.

2.3 Procedure

The experiment was delivered to subjects by computer. Subjects were presented with each passage and its associated questions simultaneously and subjects were allowed to work at their own pace. The interface of the delivery software allowed them to navigate between passages and take breaks as needed. Although other designs were explored, this sort of "power test" is designed to maximize a subject's comprehension abilities, without placing extra demands on short term memory.

2.4 Predictions

We expected that if HT is a better translation than MT, then HT should enable more accurate answers for the questions than MT, and the amount of time needed for an item in its HT condition should be shorter on average (with the exception of any bail-out effects that MT might have if its quality is too poor). The reverse patterns of results would be expected if MT is unexpectedly "better" than HT. If HT and MT are of similar quality, no differences in our measures would be expected.

2.5 Analysis Method

Each subject's test results were scored using DLI's standard for comprehension assessment against predefined scoring protocols for each item. This same scoring process was applied to MT and HT (the control case) items with the scorers blind to condition and scores. Similarly, subject reading times for test passages were normalized for length and individual subject reading rates and

against controls within each difficulty level. Intuitively, our measure of a test MT system's performance is the degree to which its output enables a subject to execute the MT-DLPT at a performance level near that of our control. With increasing difficulty we expect that the performance of a MT will degrade, thus lowering a given subject's ability to answer questions and read fluidly.

3. Experimental Results

3.1 Comprehension Accuracy

We ran 84 subjects from MIT and the surrounding community. These subjects answered questions from each of the three levels in different conditions (either MT or HT, selected at random and balanced per level). Subjects were allowed to work at their own pace, taking breaks on an individual basis. Additionally, subjects were required to take a mandatory break between sections to alleviate fatigue. Subjects took a short multiple-choice screening section, designed to predict their ability to perform in English at Level 3. Results from subjects scoring 75% or less on these screens were removed. Table 2 shows the number of trials in each condition.

Table 2 DLPT Question/Answer Trials*

	MT	HT	TOTAL
Level 1	1,402	1,403	2,805
Level 2	1,351	1,356	2,707
Level 3	1,218	1,220	2,438
TOTAL	3,971	3,979	7,950

Figure 2 shows the accuracy results from the 69 subjects who passed the screening test, where the results are pooled across subjects. Using the 70% criterion, we find that, as a whole, our subject pool passes all three levels only in the HT (human reference) condition. As the figure shows, MT has lower performance than HT.

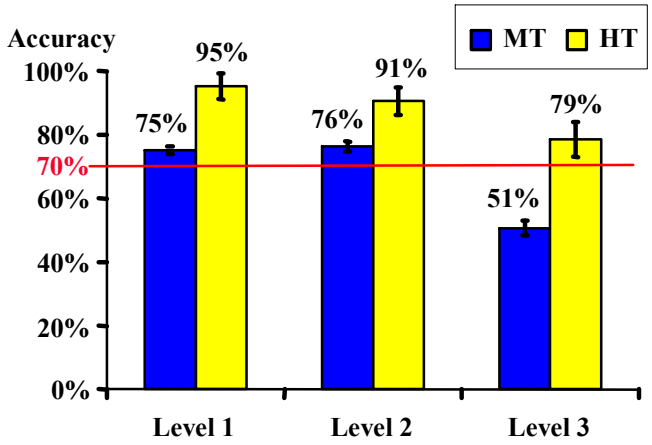

Figure 2: Performance vs. ILR Text Difficulty Level

Figure 3 shows the number of subjects who surpass the standard threshold criterion in each condition at each level. In this case, the accuracy results are not pooled across subjects. Rather, each subject must answer 70% or more of the questions correctly to pass as an individual. All 69 subjects passed Level 1 for HT and 56 passed for MT. One subject did not pass Level 2 for HT, leaving 68 passing, with a sharper drop-off to 54 subjects passing for MT. Level 3 questions are hardest, and here MT sharply degrades in performance. What is perhaps most remarkable is that 12 of the subjects managed to pass the test at Level 3 in the MT condition.

It is worthwhile to point out that reading and answering questions for Level 3 text is markedly more difficult than for Levels 1 and 2, even in the high quality human reference translations (HT). Our assessment is that these texts are college-level texts, comparable to those used in GRE reading comprehension passages. Given the difficulty of the texts, we do not expect perfect performance in the HT condition.

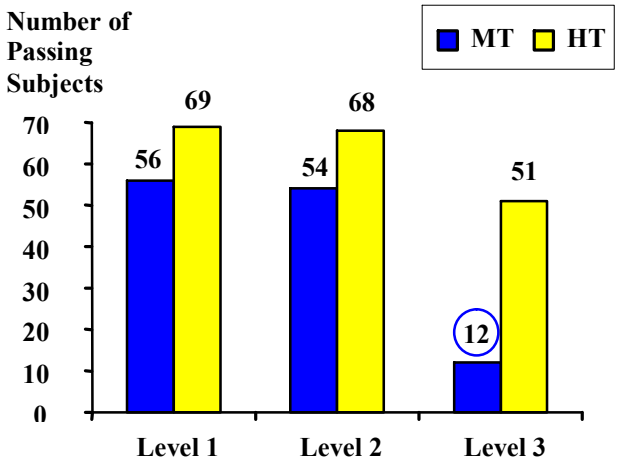

Figure 3: Number of Subjects Passing Each Level

3.2 Reading Times

In addition to question accuracy, we instrumented our experiment delivery software to keep track of the amount of time subjects spent per passage. We hypothesize that more difficult texts require more time to read (after normalizing for length). In this context, a text may be difficult if it is complicated (i.e. high ILR level) or if it is degraded in some way (i.e. incoherent, disfluent).

Anecdotally, many subjects reported using different strategies for Level 1 texts. These subjects reported that they were able to read the questions and then merely scan for the appropriate answers, given the short length of these passages. If this is true in a large number of cases, we might expect a linear model to be less appropriate for analyzing our data. We also performed an analysis of using document-level length normalization. Figure 4 shows length-normalized reading times for each level and for each condition. For each level, we see significant

differences between MT and HT conditions suggesting that MT is generally more difficult to read and/or that questions are more difficult to answer in this condition. Overall, the MT condition required 27% more time than the reference condition.

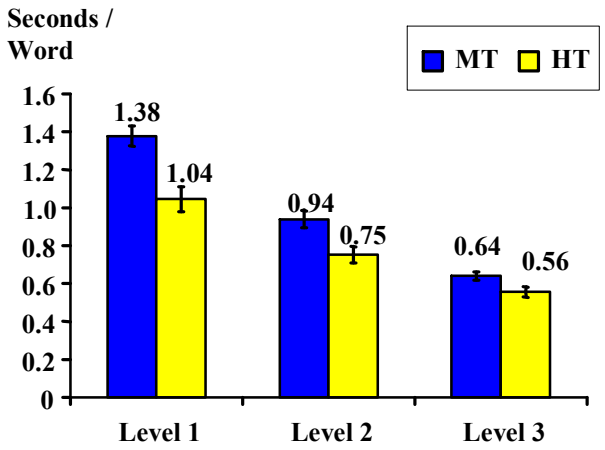

Figure 4: Length Normalized Reading Times

Across levels, subjects report different strategies: for the longer, more difficult Level 3 items, they typically look at the questions first and then scan for answers. For the shorter Level 2 items, they often read the passage first. The different ranges of reaction times reflect this type of variation in performing the task.

3.3 BLEU scores

In order to achieve the highest quality of human translation possible, the best segments from each of our four translations were used to construct the HT condition. Once selected, these segments were edited to ensure proper flow within each document. This process was done blindly with respect to the questions used in this study. Since the HT condition drew segments from any one of the four human translations, we generated three reference translations (from a set of four) excluding any segments that were used to create HT. These references were then used to score all two contrasting conditions (MT and HT).

Figure 5 shows BLEU scores for each system computed from document pools at each level. The performance of the MT system peaks at Level 2 (0.35). This higher level of performance may be a result of more closely matched train/test conditions at this level.

It is interesting to note that the performance of MT as measured by BLEU does not degrade in the same way as question-answering performance. In the case of question-accuracy, both relative and absolute accuracy was lower at Level 3 (and not passing), while the same BLEU performance at Level 1 enabled a passing level of performance. Subjects noted that they were often more able to "repair" or "make sense" of Level 1 passages in the MT condition because the questions were easier. This seems to suggest that higher BLEU scores may be necessary for the comprehension of more difficult texts.

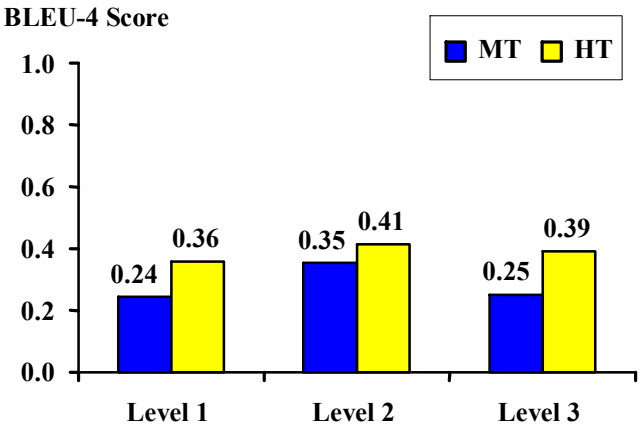

Figure 5: BLEU-4 Scores by Level

Overall we may observe the following relationships between our various factors: as quality increases from MT to HT, accuracy improves, words take fewer seconds to read, and BLEU scores improve. Each level of difficulty has its own characteristics. Level 1 and Level 3 have similar BLEU scores but Level 3 has much worse comprehension accuracy in the MT condition. Perhaps harder text must be translated with greater accuracy to be understood.

4. Conclusions and Future Work

Our study suggests that the performance of state of the art MT may enable English-native speakers to perform at ILR Level 2 in Arabic. Although today's systems may not be able to perform at Level 3, we suggest that this level of performance is a reasonable target for MT research in years to come. Our results also suggest that BLEU scores may need to be calibrated with different levels of text difficulty in order to predict effectiveness at that level. Further study would be required to bear this hypothesis out.

We view this study as an initial calibration of the ILR scale as applied to MT. Further experiments will expand the number of levels to include the intermediate "plus" levels and explore, in depth, correlation between BLEU scores and comprehension accuracy across larger document sampling pools. We may investigate other subject pools, including Government analysts, who may have greater subject matter expertise, as well as subjects with some limited Arabic knowledge. We may also want to explore whether or not these effects translate to different languages (perhaps Chinese and/or Korean). Other follow-on studies may examine the impact of reading times and cognitive fatigue that MT produces.

Acknowledgments

The main parts of this paper circulated informally on October 29, 2004 as an internal report to Charles Wayne and Joseph Olive at the Defense Advanced Projects Research Agency, whom we wish to acknowledge for their sponsorship and guidance in framing our study. The Defense Language Institute was primarily responsible for executing the study and provided the guidance and subject matter expertise needed to create and score the Arabic materials. MIT Lincoln Laboratory designed the details of the study, orchestrated the human subjects experiments, chaired an advisory committee and provided the analysis and report. We wish to thank our study's advisory committee, whose active participation throughout the study was absolutely critical: George Doddington, Ted Gibson, Kevin Knight, Mark Przybocki, and Salim Roukos. We also gratefully acknowledge the following people at the Defense Language Institute: Osaila El Khatib and Hussny Ibrahim for their Arabic expertise preparing and scoring test items. Special thanks to Rina Patel and Charlene Chuang at MIT for conducting experiments at MIT and to John Tardelli and his colleagues for conducting experiments at ARCON.

References

Child, James R., Ray T. Clifford and Pardee Lowe, Jr. 1993. "Proficiency and Performance in Language Testing". Applied Language Learning, Vol. 4.

Clifford, R, et al. 2004. "The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean". Proceedings of Language Resources and Evaluation Conference, Lisbon.

Defense Language Institute Course Catalog. 2004. ILR Skill Levels: http://www.monterey.army.mil/atfl/daa/skill.htm

Doddington, G. 2003. "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". NIST MT Web Site, February 2003. http://www.nist.gov/speech/tests/mt/

Jones, Douglas A. et al. 2005. Measuring Human Readability of Machine Generated Text: Three Case Studies in Speech Recognition and Machine Translation". To appear in Proceedings of HLT Special Session, ICASSP 2005, Philadelphia.

Lowe, Pardee. 1999. "James R. Child's Text Modes and Their Derivatives: A Compilation of Description", manuscript from course materials.

Papineni, Kishore, et al. 2001. "Bleu: a Method for Automatic Evaluation of Machine Translation" IBM Computer Science Research Report RC22176 (W0109-022) 9/17/2001 http://domino.watson.ibm.com/library/

Wayne, Charles, 2004. TIDES Program Review. Presentation at NIST Machine Translation MT-04 Workshop, June 22, 2004.

Appendix

Sample Level 2 MT-DLPT Item

Human Translation

[1] Cairo: "The Middle East" [2] Egyptian security forces at Cairo Airport prevented the retired actress Sabreen from traveling to the Kingdom of Saudi Arabia with her husband Tarek Galal Mahmoud. [3] Sabreen was trying to travel to Saudi Arabia the day before yesterday with her husband when Egyptian authorities discovered that the actress, whose full name is Sabreen Yaseen Mahmoud Abdullah and who resides in Old Cairo's Elmanial quarter, is among those barred from leaving the country. [4] The newspaper Elshark Elawsat was not able to elicit any response from the actress. The paper tried contacting her several times on her cell phone, which seemed to have been turned off, as well as on her home phone, but she was unavailable. [5] The actress Sabreen retired about four years ago after her last successful series, "OM Kolthoom," without providing any reasons for her retirement. [6] A few months ago, she got married for the second time. Her current husband, Tarek Galal Mahmoud, is the cousin of retired actress Soheir Ramzy. [7] Sabreen has one son from her previous marriage with businessman Yaser Abdulatif.

Machine Translation

[1] Cairo: "The Middle East : [2] Egyptian security forces prevented at Cairo airport Steadfasts Mu'tazilites Egyptian representative to travel to Saudi Arabia, accompanied by her husband, Jalal Tariq Mahmoud. [3] The Steadfasts Knut may travel to Saudi Arabia the day before yesterday when accompanied by her husband to the Egyptian authorities that represented, called Steadfasts Yassin Mahmoud Abdullah (residing in the area of ancient Egypt), has been included in the lists of banned from leaving the country. [4] It was not possible to obtain a response from Endurings artist has tried to "the Middle East" contacted the mobile telephone several times it was closed, and home telephone but also non-existent. [5] The representative Endurings representation for nearly four years following the introduction of the last successful "or a" without giving any reasons for the accord. [6] It was married a few months ago for the second time from her husband, Jalal current Tariq ibn Mahmoud regard represented a symbolic Suhayr withdrawn. [7] Single child from her former husband businessman Yasser Abdel-Latif.

Question:

1. Why did Sabreen have to change her travel plans?
2. What reason did Sabreen give for retiring from TV show business?
3. What is the family relationship between her and Suhair Ramzy?

Answer:

1. Security Forces stopped her from leaving Egypt.
2. No reason given.
3. Her husband's cousin