Towards an Automated Evaluation of an Embedded MT System

J. Laoudi[#,*]
[#]ARTI
Alexandria, Virginia
jlaoudi@arl.army.mil

C. Tate[±,*]
[±]Dept. of Mathematics
U. of Maryland,
College Park, Maryland
ctate@math.umd.edu

C. R. Voss[*]
[*]Multilingual Computing Group
Army Research Lab
Adelphi, Maryland
voss@arl.army.mil

Abstract. How can recent advances in automating the evaluation of machine translation (MT) engines be applied to automate the evaluation of more complex *embedded MT systems*? In this paper, we describe initial evaluation testing of FALCon, an embedded MT system where hard-copy documents are scanned into bit-mapped images, converted into online text files via optical character recognition (OCR) software, and then translated by an MT engine from one natural language into another. Our challenge with the ongoing support of FALCon systems is to evaluate when the replacement of a particular component with a new or upgraded product will yield significant improvements over the baseline system performance. In this paper we address the following questions: (i) how can we automate the baseline end-to-end evaluation of this system? and (ii) what is the relation between the accuracy of the individual components and their end-to-end accuracy, that could be used to model the system performance?

1. Introduction

How can recent advances in automating the evaluation of machine translation (MT) engines[1] be applied to automate the evaluation of more complex *embedded MT systems*[2]? Consider, for example, FALCon, an embedded system whose process flow appears in Figure 1. First the hard-copy document pages are scanned into bit-mapped images, then converted into online text files via optical character recognition (OCR) software, and then translated by an MT engine from one natural language into another [3,4]. Our challenge with the ongoing support of FALCon systems in the field is to evaluate when the replacement of a particular component with a new or upgraded product will yield significant improvements in system performance. In this paper we address the following questions: (i) how can we automate the baseline end-to-end evaluation of this system? and (ii) what is the relation between the accuracy of the individual components and their end-to-end accuracy, that could be used to model the system performance, to assess the added value of engineering new components into the system?

Our work in evaluating FALCon has led us to pursue both the task-based approach of testing users in hands-on experiments [13] and the text-based approach of assessing the system output against human reference translations [14]. The former yields *measures of effectiveness* that provide our users with an estimate of their level of accuracy on specific task applications using this system, while the latter yields *measures of performance* that capture different aspects of the system's lexical, morphological and grammatical coverage. A full system evaluation ultimately requires both types of measures and determining the relation between them.[3]

[1] See [2],[7],[8],[9]

[2] *Embedded MT systems* are defined as computational systems with one or more MT engines embedded as modules in the process flow of the system [15,16,17].

[3] The relation between these two types of measures is discussed in [12]. The insight that there are "good applications for crummy MT" [1] is another way of saying that MT engines can score highly on measures of effectiveness while scoring much less well on measures of performance. Experiments with a simple gisting MT engine have shown this [10].

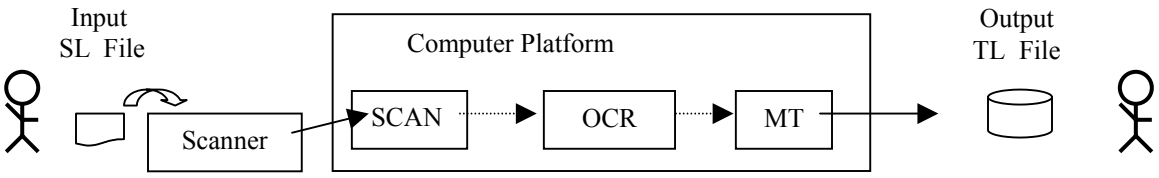


Figure 1. FALCon, an embedded MT system

For the first phase of our work on FALCon evaluation, we established a "lower bound" filtering task that English-speaking users could not effectively carry out on the source language (SL) text in Arabic once translated into English. We then tracked the content and accuracy of the text flow through the different modules of the system on that task to set a "lexical accuracy" cutoff, a minimum proportion of translated words needed for the filtering task that were semantically correct, open class, and domain-relevant, as shown in Figure 2 for the general case and, as shown in Figure 3, for the specific case on a paragraph of text translated by an Arabic-English FALCon in 1999 [16].

The stepwise tracking process was initially performed manually by the translator who evaluated the system. This was highly accurate, readily verifiable and replicable, but time-consuming. Although the process was carried out by hand, all but one of the steps were in principle easy to automate after creating a list of closed class words and another list of the domain vocabulary. The exception was the one step where the translator needed to determine which open class words generated by the MT engine failed to correspond to a content word or phrase from the SL input text (see the stick figure in the leftmost processes column of Figure 2).

The data for this evaluation process consisted of online SL texts that served as the "ground truth" (GT) input and hard-copy versions generated from these text. The GT texts were run through the MT component only,while high-quality, printed copies were scanned and run through the OCR and then MT components. In Figure 3, the results of the GT/MT and the Scan/OCR/MT processed texts are presented side-by-side for comparison. For example, it is easy to see that the OCR-ing introduced text errors because, even though the MT generates roughly the same number of TL words in both passes (fourth circle down from the top, 70 versus 68), the proportion of incorrect TL words in the OCR-ed pass is more than three times the number in the GT/MT pass (20 versus 6).

2. Approach

2.1 Multiple Component Evaluations

In 2002, two new MT engines and two different OCR software packages became available to us for testing on FALCon. For comparison here, as shown in Figure 4, we ran the same text from the 1999 test on six passes: through the Scan/OCR/MT passes on the four combinations of the two MT engines and two OCR packages, as well as through the GT/MT passes with the online data for the two MT engines. The leftmost column with the process diagram in Figure 4 indicates the automated method used for each computed items. Those marked as *auto count* were based on a simple programmed typographical word count (where "words" are strings surrounded by blanks), such as the number of "words" in the OCR output and in the MT output. The *auto inferred* items were derived by subtracting the auto count of the item at the same level from the count in the item immediately above. The *auto estimate* was inspired by the approach taken by Papineni et al., applying a "reduced" 1-gram variant of it where only the open class words in the MT output were matched against only the open class words in the five reference translations to estimate the number of semantically related open class words [14].

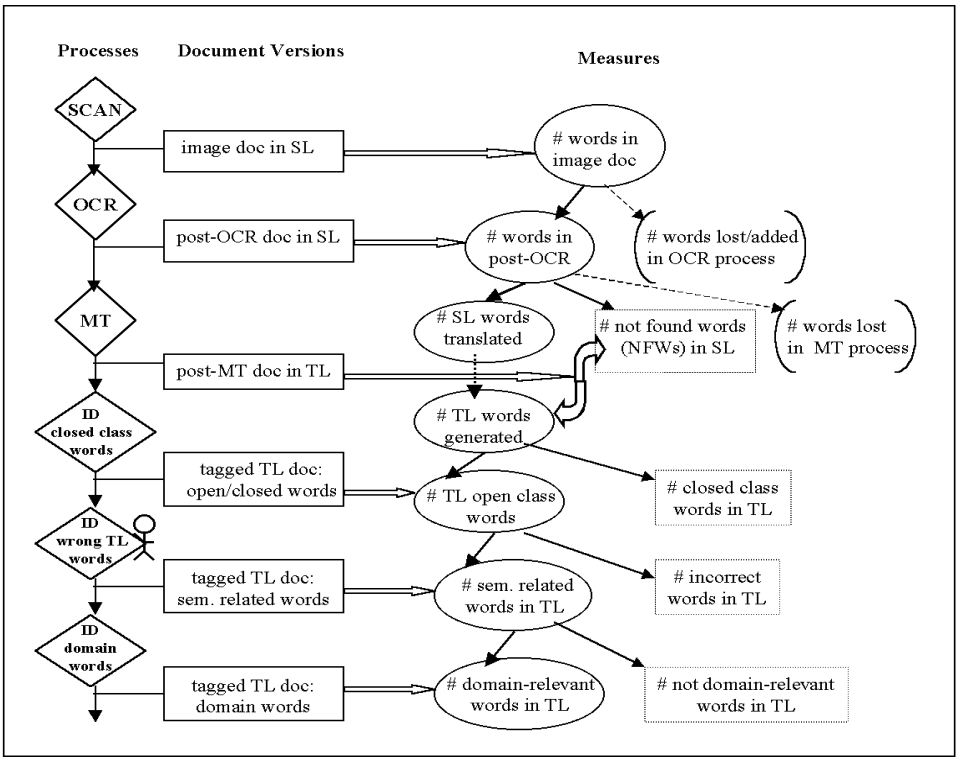


Figure 2. Tracking the Content and Accuracy of Text Across Processing Phases of FALCon

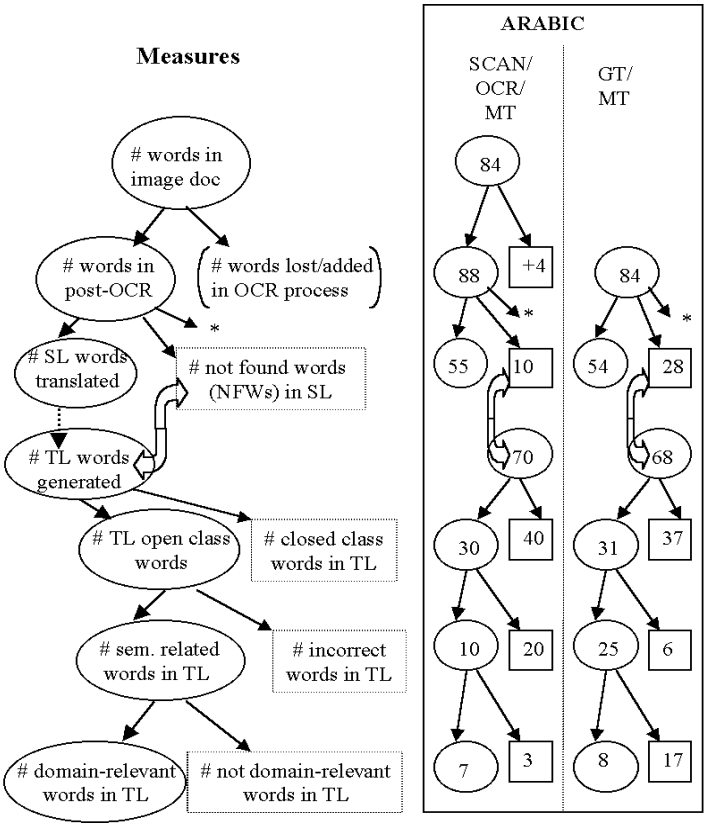


Figure 3. Two passes of processing the same text on Arabic-English FALCon in 1999

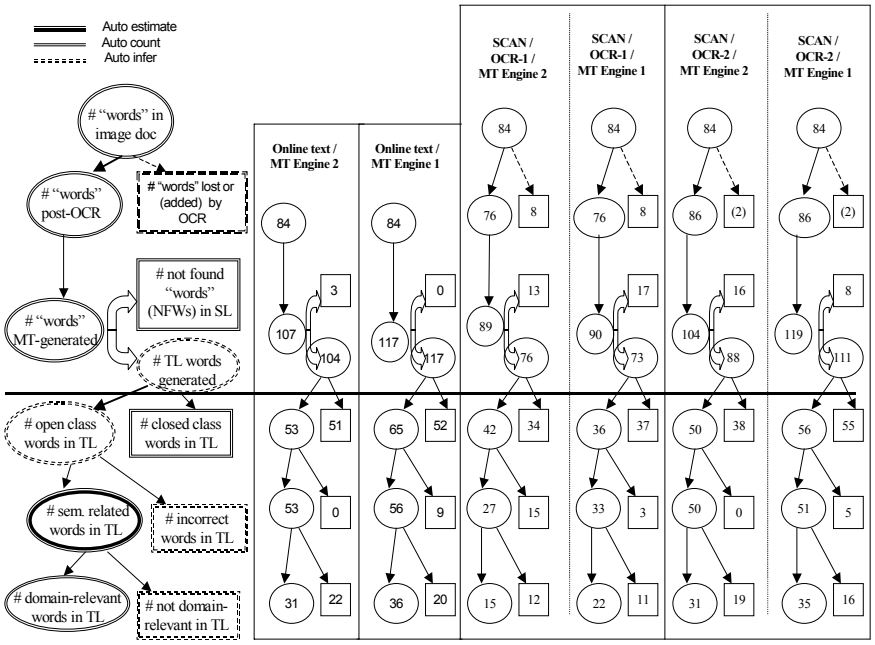


Figure 4. Results of four FALCon 2002 test variants (MT engines 1 & 2, OCR products 1 & 2)

The results in Figure 4 show that an ordering of the performance of the MT components and the OCR-MT component sequences based on the number of semantically related words in the output or target language (TL) is, with but one exception, identical to that for the number of domain-relevant words in the TL at the end of the passes:

GT-MT1 > GT-MT2 > OCR2-MT1 > OCR2-MT2 > OCR1-MT1 > OCR1-MT2

56	>	53	>	51	>	50	>	33	>	27
36	>	31	<	35	>	31	>	22	>	15

Using the same files from the automated process tracking evaluation (both GT /MT-ed and OCR/MT-ed), we also ran the Bleu 1-gram, 2-gram, 3-gram, and 4-gram scoring metric with the 5 human reference translations. The results appear in Figure 5, with (i)the GT/MT scores consistently higher than the corresponding OCR/MT scores for the same MT engine (as expected), (ii) the engine MT-1 scoring higher than MT-2, and (iii) the same ordering of sequences of components appears as in the automated tracking with the reduced Bleu 1-gram across on the matched open class words.

GT-MT1 > GT-MT2 > OCR2-MT1 > OCR2-MT2 > OCR1-MT1 > OCR1-MT2

In short, the automated, step-wise tracking and all the n-gram scores on Bleu are consistent with each other in ranking the variant-component end-to-end systems.

A few other observations are clear from this data as well. When the OCR software is weak, as is the case for OCR1, then there is little to no difference between the OCR1-MT1 and OCR1-MT2 scores because the noise level of the OCR pre-empts the two MT engines for exhibiting their strengths. In contrast, OCR2 is adequate enough that we do see differences in the OCR2-MT1 and OCR2-MT2 scores, reflecting what is already noted, that MT1 is stronger than MT2.

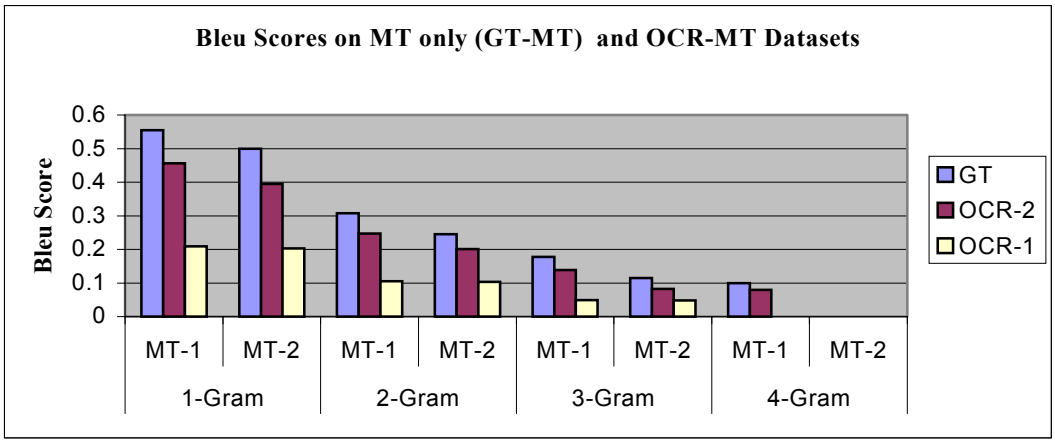


Figure 5. Bleu 1-,2-,3-, and 4-gram Scores on the same files as in Figure 4

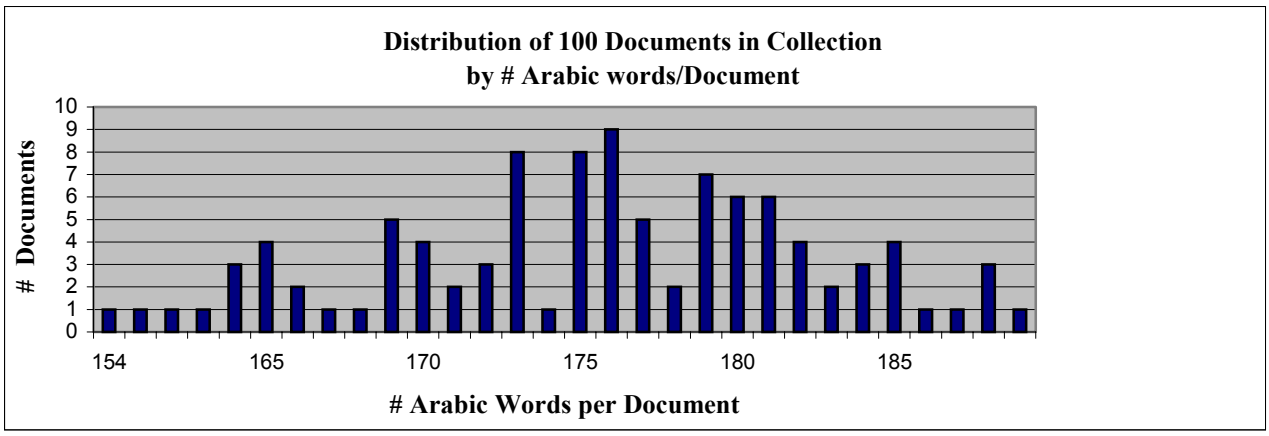


Figure 6. Distribution of documents by length

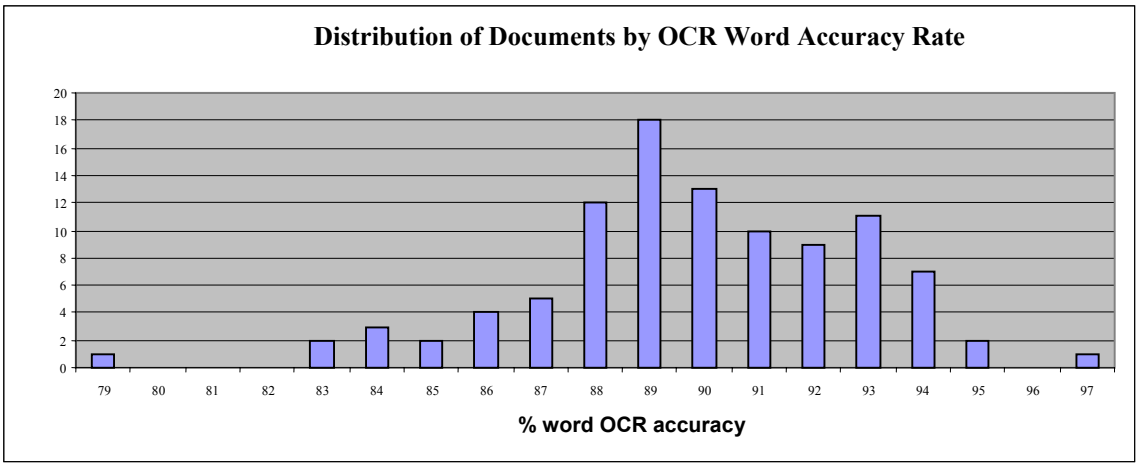


Figure 7. Distribution of documents by level of word OCR accuracy

2.2 Scaling Up the Test Dataset

Having established that the automated tracking scores and the Bleu scores present evidence for the same ranking of possible FALCon system configurations, the next challenge has been to establish that these results hold for a large dataset. We opted to leverage a 100-document collection created elsewhere that included with it numerous reference translations. Figure 6 displays the range of lengths of these documents, from 154 to 189 Arabic (typographical) words.

Working now with one MT engine and one OCR software package on the FALCon system to be evaluated with the 100 documents, we ran the hard copy versions of the documents[4] through the OCR and computed the word accuracy rate in percentage correct using the scoring algorithm from University of Nevada at Las Vegas (UNLV) algorithm, with the results shown in Figure 7. While word accuracy rates on high quality printed English texts has been reported at over 99% [5], the word accuracy rates for hard copy Arabic-script texts with comparable quality have been reported in the mid-80 to mid-90 percent range [6]. Our documents fall within this latter range.

3. Results

The online versions of the text files were run only through the MT engine and the OCR-ed files (whose scores are in Figure 7) were then run through the MT engine. These two sets of files, "GT/MT" and "OCR/MT," were then scored with the Bleu 4-gram metric using four human reference translations. Figure 8 shows the distributions of the two sets of files, with OCR/MT files scoring at the lower end and GT/MT files at the higher end of the Bleu range from 0 to 0.16, but with a fair amount of overlap in scores at the lower end.

The overlap in Bleu 4-gram scores is also evident in the box-and-whisker plots in Figure 9, where each rectangle represents the middle 50% of the documents, the median is denoted by the line inside the box, and the 25th and 75th quartiles are marked by the bottom and top of the rectangle, respectively. The whiskers extend on dashed lines to the minimum and maximum set values but no more than 1.5 time the length of the inner quartiles. Outliers beyond the whiskers show as circles [11].

Two other metrics were also run on the two sets of documents to examine whether they might be more sensitive to differences in the GT/MT and OCR/MT files' text properties. The reduced Bleu 1-gram was defined to be comparable to Figure 4's number of semantically related open class word count, using the Bleu 1-gram on only the open class words [14]. Figure 9 shows that the reduced Bleu 1-gram and the f-score [8] yield a clearer distinction between the two sets.[5] Given that, on a pairwise basis, the MT score for a file is always higher than its OCR/MT version, we now address one of the questions from the introduction, concerning the relationship between the accuracy of the components and their sequencing:

> *When OCR word accuracy rates go up, do OCR/MT accuracy rate in the system also go up, as measured by the Bleu 4-gm, the reduced Bleu 1-gm,or the f-score?*

To answer this question, we tested with the Pearson product-moment correlation coefficient (r), also called the correlation coefficient, to measure the linear relations between the two variables. Figure 10 displays the scatterplots for these three correlation analyses and the table below it show the correlation results: there is only very weak evidence for answering *yes* to this question, given these metrics and this dataset.

[4] We created high-quality, hard copy versions by printing the original online texts with a laser printer, giving us an upper bound for the OCR results.

[5] All text was converted to lower case before running the metrics.

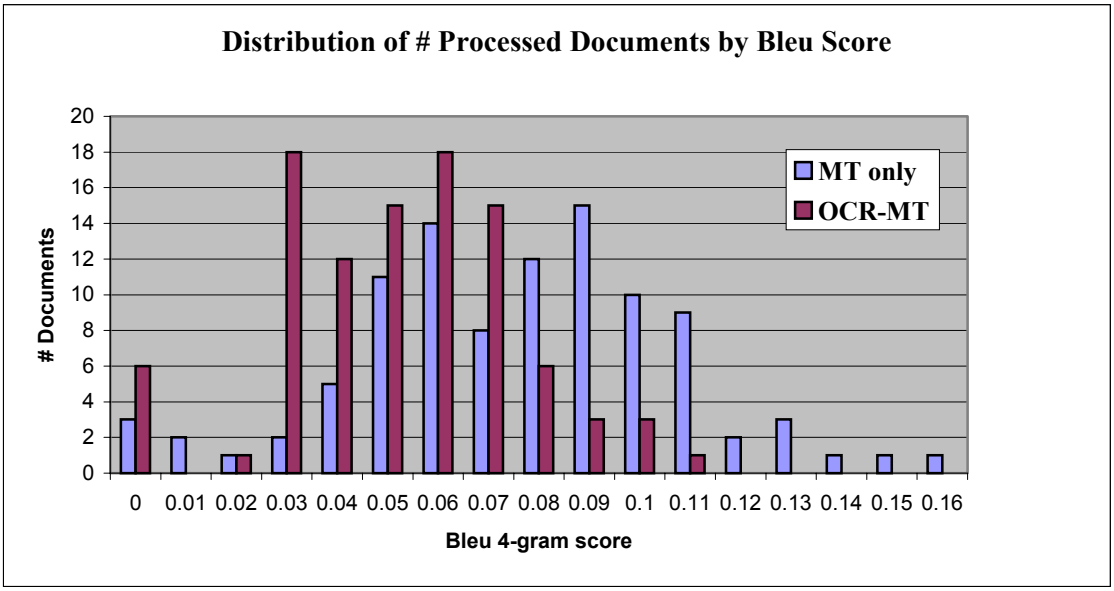


Figure 8

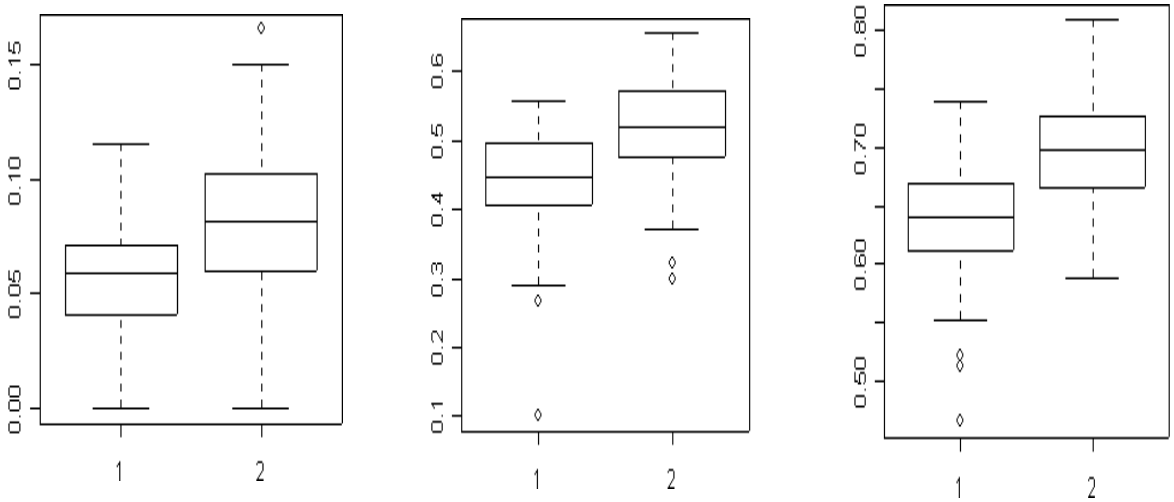
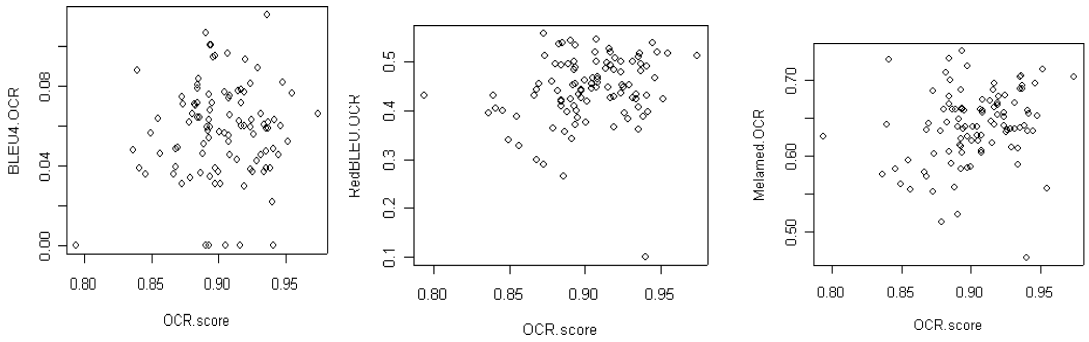


Figure 9. Three types of evaluation scores on 100 Arabic-to-English translated files
1 = hard copy files scanned, OCRed, and-MTed **2** = online files MTed only



Bleu 4gram & OCR Reduced Bleu 1gram & OCR f-score & OCR

Figure 10. Scatterplot of MT & OCR scores for the same 100 Arabic-English translated files

Table 1. Results of Correlation Testing on Scores in Figure 10

MT score with OCR word accuracy score	**Correlation r**	**p-value**	**Significant at .05 level?**
Bleu 4-gm	.09	.34	No
Reduced Bleu 1-gm	.19	.052	Yes
f-score	.24	.01	Yes

The results of the testing show that the Bleu-4-gram metric has almost no correlation (r is close to zero) between the OCR and the OCR/MT scores on this dataset. The Bleu metric appears thus not to provide enough sensitivity for our datasets. When applied to compare GT/MT and OCR/MT documents in Figure 9, the 4-gram scores for the two sets overlapped by almost a quartile, while the reduced Bleu 1-gram and the f-score results overlapped minimally. Also, when used in the correlation testing of OCR and OCR/MT documents in Table 1, even given the 100 source document set size, there is no statistically significant result detected.

In distinct contrast with this result,
(i) the correlation between the ***f-score*** on the OCR/MT texts & the ***word accuracy score*** on the OCR-ed texts and
(ii) the correlation between the ***reduced Bleu 1-gram*** scores on the OCR/MT texts and the ***word accuracy score*** on the OCR-ed texts, *though both yield relatively weak correlations, they are also significant statistically*.

One possible interpretation for the low correlations in these results is that the OCR accuracy measure, as it is currently applied to all words equally, may be too broad or at the wrong granularity for MT evaluation. Perhaps limiting the OCR score to an open-class word accuracy would correlate better with the reduced Bleu 1-gram score, as relevant to the domain-filtering task for which FALCon was developed.

Alternatively, we may be missing a hidden factor in the analyses. The correlation might be enhanced by controlling for a variable such as the length of the document segments in Arabic words. From Figure 11, we can see the wide distribution of segment lengths in our test collection. Since MT on longer segments tends to be more difficult than MT on shorter segments (with or without OCR noise), the extra long segments in our collection may mask a stronger correlation by biasing the results with test data that makes all the MT engines score poorly. One next step in our research will be to break out the data by segment length and test for OCR/MT score correlations with OCR word accuracy rate on segments of different lengths.

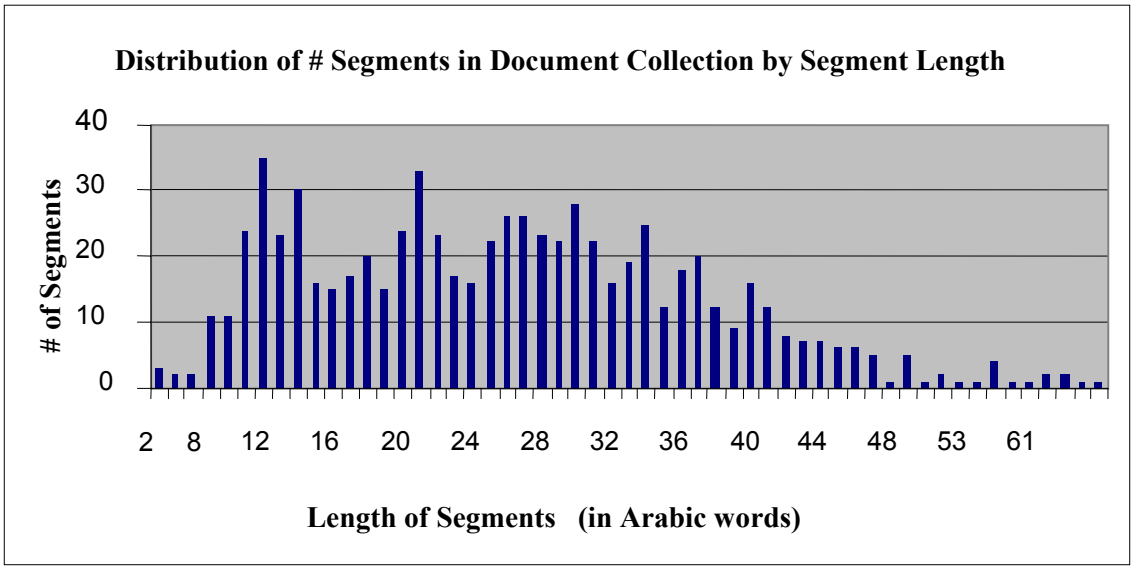


Figure 11

4. Conclusion and Future Work

In this paper we have shown our initial work in applying the recent advances in the evaluation of machine translation (MT) engines to automating the evaluation of more complex *embedded MT systems*. We have established by experimentation that our automated approach adequately validates our prior manual evaluation results, where in a baseline end-to-end assessment of FALCon for a domain-filtering task, the translation of 1-gram open-class words is crucial to system performance accuracy.

One of our challenges remains the ongoing evaluation of FALCon systems for a wider range of input data types and for possible improvements with new products and upgrades. The method of evaluation presented here has been applied to other language pairs on FALCon, including Korean-English and Chinese-English. In the future, with further testing of language pairs with lower quality components and further testing of degraded documents with much lower OCR scores, the method presented here of automatically evaluating FALCon will enable us to rapidly assess the lower bounds of performance for FALCon variants on a wider range of real-world documents.

Additional research is required to model the relations between OCR accuracy rates and OCR/MT metrics, possibly involving the additional factor of segment length. Ultimately these measures of performance will need to be tied to experimental results and task-based measures of effectiveness, so that abstract performance scores can be interpreted by users in terms of tasks they do and understand.

References

[1] Church, K. and E. Hovy (1993) "Good Applications for Crummy Machine Translation." *MachineTranslation*, 8, 239 - 258.

[2] Doddington, G. (2002) "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurence Statistics." In *Proceedings of HLT 2002*, Human Language Technology Conference, San Diego, CA. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

[3] 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.

[4] Fisher, F., C. Schlesiger, L. Decrozant, R. Zuba, M. Holland, and C.R. Voss (1999) "Searching and Translating Arabic Documents on a Mobile Platform," In *Proceedings of the Advanced Information Processing and Analysis Conference* (AIPA 99), Washington, DC.

[5] Kanai, J. (1996) "Character Recognition" In *Survey of the State of the Art in Human Language Technology*, Chapter 13.10, http://cslu.cse.ogi.edu/HLTsurvey/ch13node12.html

[6] Kanungo, T., G. Marton and O. Bulbul (1998) "Performance Evaluation of Two Arabic OCR Products" In *Proceedings of AIPR Workshop on Advances in Computer Assiste Recognition*, SPIE vol. 3584.

[7] Leusch, G., N. Ueffing, and H. Ney (2003) "A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation" In *Proceedings of the MT Summit IX* , New Orleans, LA.

[8] Melamed, I.D., R. Green and J. P. Turian (2003) "Precision and Recall of Machine Translation" *Proteus Technical Report #03-004,* a revised version of the paper presented at NAACL/HLT 2003, Edmonton, Canada.

[9] Papineni, K., S. Roukos, T. Ward, and W. Zhu (2001) "Bleu: a method for automatic evaluation of machine translation", *IBM Research Report* RC22176, Yorktown Heights, NY.

[10] Resnik, P. (1997) Evaluating Multilingual Gisting of Web Pages, in *Proceedings of the AAAI Symposium on Natural Language Processing for the World Wide Web*, Stanford, CA.

[11] Rice, J.A. (1995) Mathematical Statistics and Data Analysis, Duxbury Press, Belomont, CA.

[12] Roche, J.G. and B.D. Watts (1991) Choosing Analytic Measures. *The Journal of Strategic Studies,* Volume 14, pages 165-209, June.

[13] Voss, C.R. et al. (forthcoming) Task-Based Evaluation of Machine Translation Engines. Final Project Technical Report, Center for the Advanced Study of Language (CASL), University of Maryland, College Park, MD.

[14] Voss, C. (2003) "MT Evaluation for FALCon: Issues in Component-level and End-to-End Evaluation" *DARPA TIDES MT Evaluation Meeting,* Gaithersburg, MD.

[15] Voss, C.R. and F. Reeder (1998) *Proceedings of the First Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*, Association for Machine Translation in the Americas (AMTA'98), Langhorne, PA.

[16] Voss, C.R. and C. Van Ess-Dykema (2000) "When is an Embedded MT System 'Good Enough' for Filtering?" In *Proceedings of the Second Workshop on Embedded MT Systems*, ANLP-NAACL2000, Seattle, WA.

[17] Voss, C.R. and C. Van Ess-Dykema (2002) Guest editors, Special Issue on Embedded Machine Translation Systems, *Machine Translation,* vol. 17, issues 2-4.