Exploitation of an Arabic Language Resource for MT Evaluation:

Using Buckwalter-based Lookup Tool to Augment CMU Alignment Algorithm

^Clare R. Voss, ^*Jamal Laoudi, ^Jeffrey Micher

^Army Research Laboratory *Advanced Resources Technology, Inc.

Adelphi, MD USA Alexandria, VA

E-mail: {voss, jlaoudi, jmicher | @arl.army.mil}

Abstract

Voss et al. (2006) analyzed newswire translations of three DARPA GALE Arabic-English MT systems at the segment level in terms of subjective judgment scores, automated metric scores, and correlations among these different score types. At this level of granularity, the correlations are weak. In this paper, we begin to reconcile the subjective and automated scores that underlie these correlations by explicitly "grounding" MT output with its Reference Translation (RT) ***prior to*** subjective or automated evaluation. The first two phases of our approach annotate {MT, RT} pairs with the same types of textual comparisons that subjects intuitively apply, while the third phase (not presented here) entails scoring the pairs: (i) automated calculation of "MT-RT hits" using CMU aligner from METEOR, (ii) an extension phase where our Buckwalter-based Lookup Tool serves to generate six other textual comparison categories on items in the MT output that the CMU aligner does ***not*** identify, and (iii) given the fully categorized RT & MT pair, a final adequacy score is assigned to the MT output, either by an automated metric based on weighted category counts and segment length, or by a trained human judge.

1. Introduction

Voss et al. (2006) analyzed the newswire translations of three DARPA GALE Arabic-English machine translation (MT) systems at the segment level in terms of subjective judgment scores, automated metric scores, and the correlations among these different score types. At this level of granularity, while one automated metric[1] clearly correlated better than the other automated metrics with the subjective judgment scores, overall the correlations were weak. In this paper, we begin to reconcile the subjective and automated scores that underlie these correlations by explicitly "grounding" MT output segments with their Reference Translation (RT) ***prior to*** subjective or automated evaluation

The first section of the paper introduces our approach to tackling MT evaluation at the segment level where we exploit our Buckwalter-based Lookup Tool (BBLT) to augment the "search space" of a reference translation (RT) with BBLT translations of the original source segment. The full approach consists of three stages: (i) an automated calculation of "MT-RT hits" using the CMU aligner from METEOR, followed by (ii) an extension phase where the BBLT serves to help identify six other categories of matches and non-matches on items in the MT output that the CMU aligner did ***not*** handle, and then (iii) given the fully category-annotated {RT, MT} pair, a final adequacy score is assigned to the MT output, either by an automated metric based on weighted category counts and segment length, or by a trained human judge. We describe the first two stages of our approach and the six annotated categories as they apply to the {RT, MT} pair for one Arabic MT input segment. In the Results and Ongoing Work section, we show how these two stages yield various combinations of annotation categories on the outputs of six different current Arabic-English MT engines. We conclude the paper by reviewing the weak correlation results from Voss et al. (2006) as they relate to our plans to test for correlations on subjective judgments collected on color-coded annotated {RT, MT output} pairs with various automated metrics run on these pairs.

[1] METEOR (Lavie et al. 2004; Lavie and Agarawal 2007)

2. Approach

Before describing the software and computational steps for phases (i) and (ii) of our approach, we describe the color-coded annotations that are generated during these phases to document various types of textual comparisons that subjects intuitively apply to {RT, MT output} segment pairs when scoring them for translation adequacy.

2.1 Category Annotations

The categories are described below for the omniscient annotator who is annotating text in {RT, MT output} pairs as shown in Figure 1. We expect, as in the development and application of all annotation schemes, that these category definitions will require iterative refinement after being assessed for inter-annotator reliability. In phase (iii) categories are weighted in text-based automated metric alternatives that correlate with subjective judgments.

SOURCE: وتحتاج طائرات البوينغ 737-003 الى مدرج بطول 0022 مترا على الاقل للهبوط او الاقلاع.

RT: A Boeing 737-300 requires a runway that is at least 2200 meters long for take off and landing

MT: The Boeing-737-300 aircraft to included the length of at least 2200 metres landing or take off.

Figure 1: Arabic Newswire Segment with Reference Translation (RT) and Machine Translation (MT) Output

Automated metrics such as BLEU, NIST, and METEOR identify the correct translations in terms of

(i) "exact hits" and ***"synonym/stemmed hits"*** where the MT output correctly matches the RT text. In Figure 2. below, category (i) tokens are annotated in green in the RT text and MT output, after being aligned by the CMU software, matching literally or on synonym from WordNet or by stemmed matching of literal or synonym.

But these metrics do not give credit for other types of correctly translated items in the MT output:

(ii) "RT gaps" where the MT engines correctly output text content that the human reference translator did not capture, either by mistake or by intentionally opting to omit content believed to be obvious to an English speaker. In Figure 3, category (ii) tokens are annotated in blue, as occurs with the word "aircraft" in the MT output, that is missing in the RT. The BBLT analysis identifies this inconsistency because it displays all tokens in SL with their own column, as can be seen in Table 2, second column from the left, for the "aircraft" token.

(iii) "paraphrase hits" where the MT correctly outputs content equivalent semantically to the RT, but not literally identical. In Figure 3. the RT phrase "at least 2200 meters long" corresponds semantically to the MT output phase "the length of at least 2200 metres". The non-literal, but semantic correspondence is annotated in blue in the RT and the MT output. BBLT together with WordNet can identify correspondences such the "long"/"length" in the example. [2] We expect that ultimately a source of monolingual paraphrases and alternative equivalent multi-word expressions can be added to this identification task (Ellsworth and Janin, 2007).

(iv) "RT-MT dual divergences" where the MT is literally correct, but does not match the RT term even though the MT and RT terms correspond in this context without distorting the meaning due to colloquial or idiomatic expressions. The terms are annotated in purple, for example in Figure 3 with "and" in the RT and "or" in the MT output. BBLT provides the terms for spotting these divergences.

(v) "NFW transliterations" where the MT correctly retains terms in its output for which it has no translation, typically new names. These out-of-vocabulary (OOV) terms should not be discarded by MT engines even though they are not fully correct, because they may be adequate spell-outs of names that MT user will work with. The OOV, transliterated term "Alam" in the MT output and its translated term in the RT "found out" are annotated in purple, for example in Figure 5c. The BBLT can be run with its transliteration feature on, enabling a non-Arabic reader to see transliterations aligned with their translations.

[2] WordNet defines synset with "length (a section of something that is long and narrow)"

Furthermore automated metrics do not explicitly identify two types of MT errors

(vi) *"MT gaps"* where the MT output is incorrect by failing to contain content corresponding to content word(s) in RT. These terms in the RT are annotated in yellow and do not have a corresponding term in the MT. The BBLT analysis will identify these since all content words in SL will have a column in the output and this term will not show. For example, in Figure 3. the RT verb "requires" does not have a corresponding term in the MT output. (Note that some MT systems, in being optimized for a particular automated metric, end up dropping NFWs to boost their score. This pattern can be detected by the BBLT analysis that finds RT items that match in the BBLT table, but fail to match MT items, to identify "*MT drops.*"

(vii) "MT lexical selection errors" where particular word translated is incorrect for the context. BBLT may identify these since alternate translations of a word may be in other rows of the column of the SL token and share no terms in WordNet synsets. (The BBLT analysis enables us to distinguish such errors from the "*MT hallucinations*" of statistical MTs, where the lexical selection driven by training data does not correspond to any Buckwalter or dictionary translation of any of the SL words.) These forms of incorrect terms are annotated in red in both the RT and the MT outpur. For example in Figure 3., the MT "included" is a mistranslation of the RT "runway" as can be seen in BBLT. When the mistranslations are close with some shared semantics, then the annotation is in gray as shown in Figure 5b., where the RT "found out" and the MT output "aware of" are both in gray.

2.2 Annotation Algorithm

The process for annotating the {RT, MT output} pairs starts with (i) the CMU alignment phase and then proceeds to (ii) a BBLT analysis phase.

2.2.1 CMU Alignment

We start by inputting a pair of RT and MT segments into the automatic word aligner from CMU's METEOR (also used within CMU's MEMT algorithm), for a first-pass analysis of the exact hits and synonym/stemmed hits, in category (i) above. The results of this phase for the RT and MT from Figure 1 are shown in Table 1 below.

RT	MT	CMU category
1	1	artificial
2	2	exact
3	3	exact
9	10	exact
11	12	exact
12	13	wn_synonymy
15	16	exact
16	17	exact
18	14	exact

Table 1. Results Table output by CMU Aligner on {RT,MT output} pair from Figure 1.

The numbers in Table 1. stand for word positions in the RT and MT. For example, RT 18 and MT 14 (in last row) correspond the match terms "landing". To illustrate the hits found in this way, the corresponding items in the RT and MT segments are annotated in green in Figure 2. We can also see that token 12 for "meters" in the RT column of Table 1 matches token 13 in the MT for "metres:" the algorithm reconciles the different spellings via a WordNet synonymy check.

RT: A Boeing 737-300 requires a runway that is *at least* 2200 meters long for take off and landing

MT: The Boeing 737-300 aircraft to included the length of *at least* 2200 metres landing or take off

Figure 2. Annotated {RT, MT output} Pair from Fig. 1 following processing step by CMU aligner

After the alignment is run, new columns are added into the CMU results table and their contents are filled as follows: For each RT position, add in the RT word as second column. For each MT position, add in the MT word as fourth column. Keep the CMU results as generated by the aligner, now in the fifth column. For the exact matches and WordNet synonym/stem matches, the sixth, seventh, and eight columns are filled with "accept", blank, and "MT correct". For the "artificial" matches, the sixth column is marked "Review", since a human needs to compare the RT and MT items of that row for scoring. Typically the "artificial" matches are pairs of closed class words that are not translations of each other. We allow the human reviewer to assign partial credit if it is clear that words correspond to each other, as in the "a" and "the" in the given example. Only the human review and credit allocation in the sixth column of "artificial" rows need be done manually.

2.2.2 BBLT Analysis

In the second phase, the source language sentence is input to the BBLT (available both as web service and as GUI) and a GUI table result appears, see Table 2 for the source segment in Figure 1. The analysis that follows extends the matched alignment that occurs between the RT and BBLT, and between the MT and BBLT. The BBLT Results Screen shows the English meanings in a table where each column corresponds to an Arabic token in the input sequence, but presented in reverse order. That is, the right-to-left Arabic order of the original input sequence is transformed in the Results Screen table into a left-to-right order.

For each such "BBLT match alignment" of the CMU non-matched words in the RT or the MT for which there is also a *corresponding* column in the BBLT GUI table, the Results Table is augmented with a new row. Here, *corresponding* refers to some shared text content that can be automatically identified in three cases: the RT word matches word in BBLT column, MT word matches word in BBLT column, or both (where the RT word and the MT word match distinct words in the same BBLT column, i.e., in different rows of that column.)

Each new row is binned into one of the categories (ii)-(vii) identified above. The algorithm for filling the first/second and third/fourth column pairs of these new rows is based on content inspection of the corresponding BBLT column. For example, the word "aircraft" shows up in BBLT as well as in position 4 of the MT, but no equivalent is present in the RT, so the first//second RT pair is left blank and the third/fourth pair is filled with "4" and "aircraft". This is categorized as ***(ii) RT gap*** and colored blue, since the word is a correct translation but the human reference translator opted not to include it. The ***(iii) MT paraphrase*** case is illustrated in augmented CMU+BBLT Table 3 in the RT "13 long" and the MT pair "8 length".

RT: A Boeing 737-300 requires a runway that is *at least* 2200 meters long for take off and landing.

MT: The Boeing-737-300 aircraft to included the length of *at least* 2200 metres landing or take off.

Figure 3. Annotated {RT, MT output} Pair from Fig. 2 following processing steps with BBLT analyzer

3. Results and Ongoing Work

We now show how these two stages result in a range of different categories on the outputs of six different current Arabic-English MT engines. Figure 4. presents source language segment, a reference translation, and then the machine translation outputs for that same input segment. Figures 5a through 5f show the pairwise color-coded annotation of the MT and RT pairs.

While all the MTs translated the subject of the sentence correctly, only MT 1 is successful in situating the full subject NP at the front of the sentence. MT2 selected a partially correct translation for the leading verb and MT5 found the correct translation, but both left it in sentence-initial position. MT3 transliterated the leading verb and MT 6 mistranslated it, and again both also left it in sentence-initial position. MT4 found the verb but appears to have moved the sentence-final termporal expression to the front of the sentence, leaving the verb-subject order unchanged. Given the preponderance of verb-initial sentences in Arabic, it is quite surprising that only one MT engine handled this construction correctly.

Similarly while all the MTs indicate a start date of a battle and a time of Wednesday, only MT 1 and MT 4 are successful in moving the time out of sentence-final position to get the correct verb-event reading where the time modifies the knowing/finding out, not noun phrase-event of the start date of the battle. The sequence of RT-MT shared color coding for adequacy (recall that green and blue indicate correct matches and gray indicates a partial match) and the fluency of the text within a singly-colored sequence indicate that MT 1 should be subjectively judged the best translation and MT 6 the worst (recall that red is error and yellow is missed terms).

Given the ease with which we can "see" and rank MT outputs for their translation adequacy with this color-coded annotation, the next challenge in our phase (iii) research is to identify the set of annotated textual comparisons that subjects use in judging annotated MT output so that these can be incorporated into automated evaluation metrics. We will know that we have made progress in reconciling the subjective and automated scores when we can revisit the translations from the scatterplots in Figure 6 (from Voss et al. 2006) and show that these weak correlations can be improved with annotated text comparisons relevant both to subjects judging adequacy and to MT developers in need of sensitive, well-calibrated automated metrics for training and optimizing their MT engines.

References

Buckwalter Arabic Morphological Analyzer (BAMA), Version 2.0, LDC Catalog number LDC2004L02, www.ldc.upenn.edu/Catalog

Ellsworth, M. and A. Janin (2007) "Mutaphrase: Paraphrasing with FrameNet." In Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic.

Habash, Nizar. (2004) Aragen: Large Scale Arabic Morphological Generation. Poster Presentation. TECH 2004. University of Maryland College Park. March 19, 2004 http://clipdemos.umiacs.umd.edu/Aragen/

Jayaraman, S.and A. Lavie (2005) "Multi-Engine Machine Translation Guided by Explicit Word Matching". In Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT-2005), Budapest, Hungary.

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

Lavie, A. and A. Agarwal, (2007) "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments" In Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics, Prague, Czech Republic.

Strassel, Stephanie and Andrew W. Cole (2006) "Corpus Development and Publication" In Proceedings of LREC, Genoa,Italy .http://papers.ldc.upenn.edu/LREC2006/CorpusDevelopmentAndPublication.pdf

Voss, C., J. Micher, J. Laoudi, C. Tate (2006) "Ongoing Machine Translation Evaluation at ARL," Presentation, In Proceedings of the NIST Machine Translation Workshop, Washington, DC.

Zawaydeh, B. and Z. Saadi (2006) "Orthographic Variations in Arabic Corpora," Government Users Conference,www.basistech.com/knowledge-center/Arabic/orthographic-variations-in-arabic.pdf

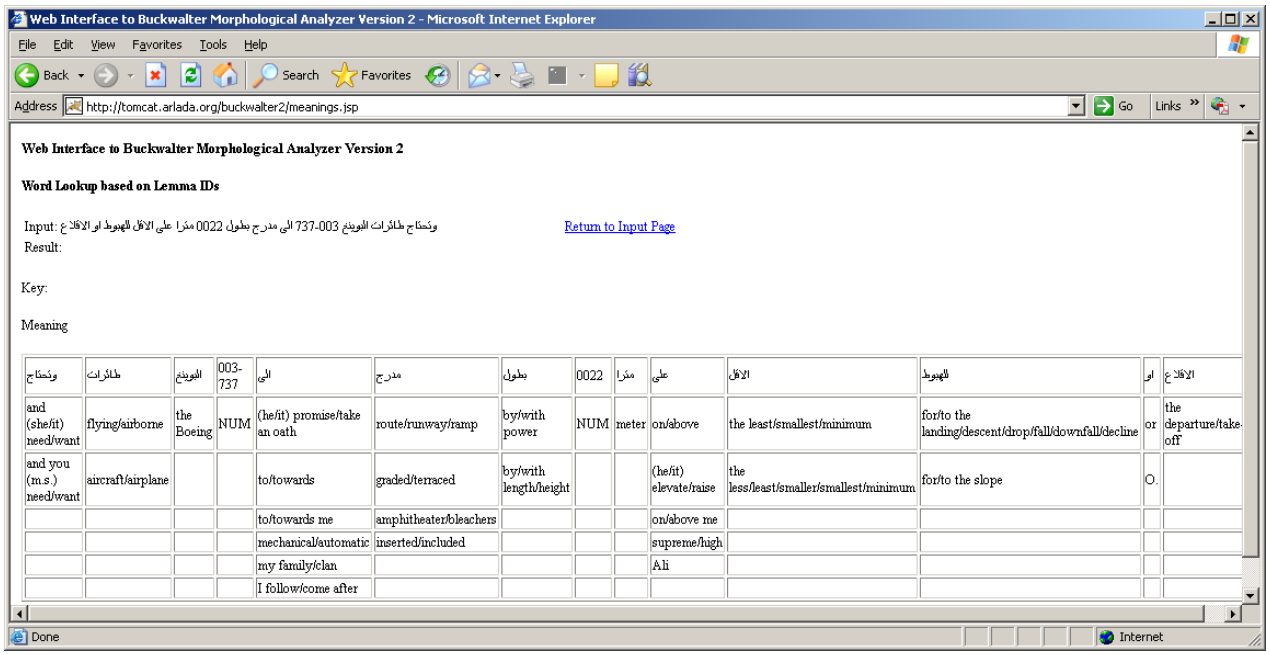


Table 2. Screenshot of Results Table from Buckwalter-Based Look-up Tool (BBLT) for SL segment in Figure 1.

RT	MT	CMU Algorithm	+ BBLT Alignment	<notes>	CMU+BBLT	category
1 A	1 The	artificial	Review:	< indef **vs**. def article>	MT partial	**vii**
2 Boeing	2 Boeing	exact	Accept		***MT correct***	**i.**
3 737-300	3 737-300	exact	Accept		***MT correct***	**i.**
	4 **aircraft**		**BBLT:**	***RT gap***	***MT correct***	**ii.**
4 requires			**BBLT:**	***RT correct***	***MT gap***	**vi**
	5 to					
5 a						
6 **runway**	6 **included**		**BBLT**:		*MT* **error**	**vii**
7 that						
8 is						
	7 the					
13 long	8 length		**BBLT:**	MT paraphrase		**iii**
	9 of					
9 at	10 at	exact	Accept		***MT correct***	**i.**
10 least	11 least	exact	Accept		***MT correct***	**i**
11 2200	12 2200	exact	Accept		***MT correct***	**i**
12 meters	13 metres	wn_synonymy	Accept		***MT correct***	**i.**
14 for						
15 take	16 take	exact	Accept		***MT correct***	**i.**
16 off	17 off	exact	Accept		***MT correct***	**i.**
17 and	15 or		**BBLT**	<idiomatic>	*MT dual diverg*	**iv**
18 landing	14 landing	exact	Accept		***MT correct***	**i**

Table 3. Results of CMU alignment (boxed) and
Results of combined CMU + BBLT matching analysis with categories

Source Language Text:
علم قائد الجيش بتاريخ ابتداء المعركة يوم الاربعاء.
Reference Translation:
The Army commander found out about the start date of the battle on Wednesday.

MT 1: the army leader knew on Wednesday in the clash beginning date..
MT 2: Aware of the army commander on the battle beginning on Wednesday.
MT 3: Alam, the army commander by the battle beginning history on Wednesday
MT 4: Day of Wednesday knew commander of the army in date of start the battle.
MT 5: know leader army with date starting the battle wednesday.
MT 6: the flag of the Army Commander on beginning the battle on Wednesday

Figure 4. Output from Six Machine Translation Engines

CMU analysis on RT & MT 1:
RT: the army commander found out about the start date of the battle on wednesday
MT 1: the army leader knew on wednesday in the clash beginning date

1 1 exact
2 2 exact
5 3 artificial
6 4 artificial
7 8 exact
8 10 wn_synonymy
9 11 exact
13 5 exact
14 6 exact

CMU + BBLT analysis on RT & MT 1 (categories i, ii, vii[3])
RT The Army commander found out about the start date of the battle on Wednesday.
MT 1: the army leader knew on Wednesday in the clash beginning date

Figure 5a. CMU analysis and CMU+BBLT analysis on on RT & MT1

CMU + BBLT analysis on RT & MT 2 (categories i, vi, vii)
RT The Army commander found out about the start date of the battle on Wednesday.
MT 2: Aware of the army commander on the battle beginning on Wednesday.

Figure 5b. CMU+BBLT analysis on on RT & MT2

CMU + BBLT analysis on RT & MT 3 (categories i, ii, v)
RT The Army commander found out about the start date of the battle on Wednesday.
MT 3: Alam, the army commander by the battle beginning history on Wednesday

Figure 5c. CMU+BBLT analysis on on RT & MT3

CMU + BBLT analysis on RT & MT 4 (categories i, ii)
RT The Army commander found out about the start date of the battle on Wednesday.
MT 4: Day of Wednesday knew commander of the army in date of start the battle.

Figure 5d. CMU+BBLT analysis on on RT & MT4

CMU + BBLT analysis on RT & MT 5 (categories i, ii)
RT The Army commander found out about the start date of the battle on Wednesday.
MT 5: know leader army with date starting the battle wednesday.

Figure 5e. CMU+BBLT analysis on on RT & MT5

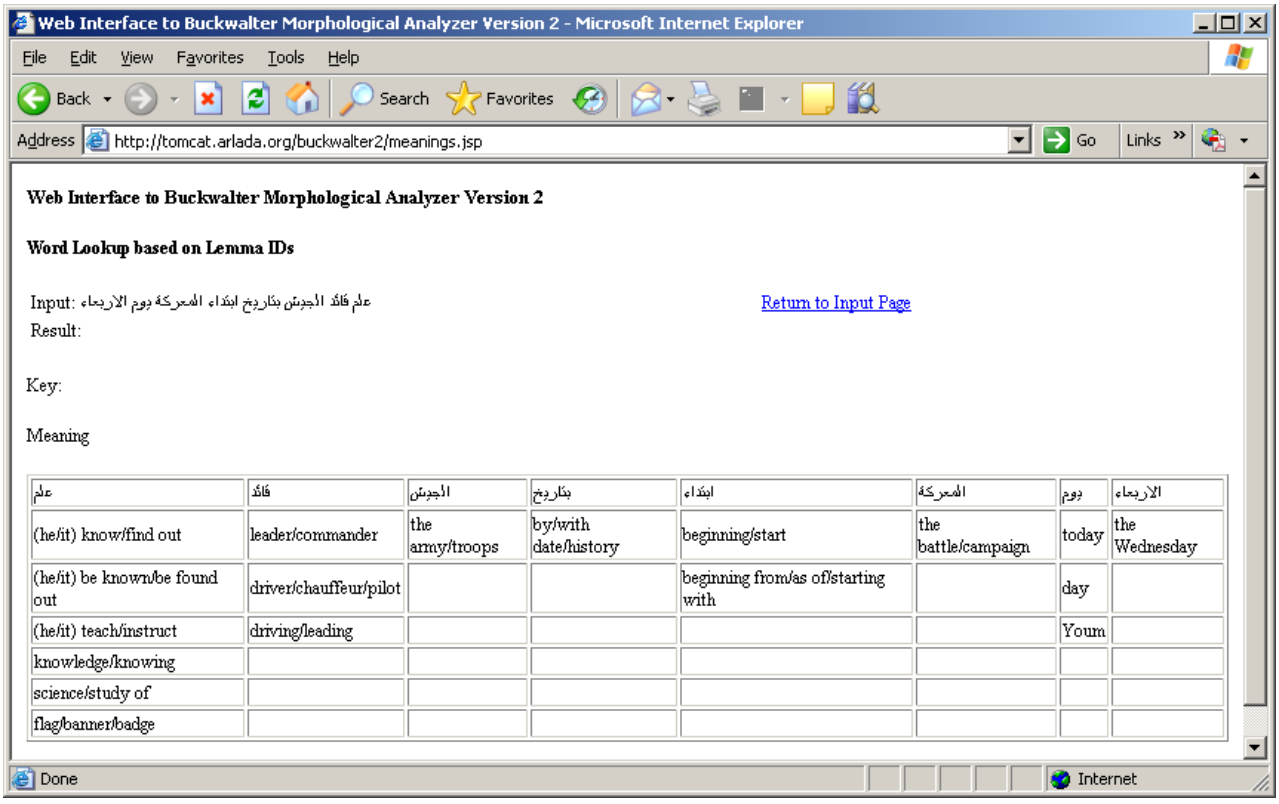
[3] "Battle" in the RT and "clash" in the MT are in gray because they would require human intervention to indicate partial credit, as the CMU engine did no match them using WordNet and the BBLT does not match "clash".

CMU analysis on RT & MT 6:
RT: the army commander found out about the start date of the battle on wednesday
MT 6: the flag of the army commander on beginning the battle on wednesday

1 1 exact
2 5 exact
3 6 exact
7 4 exact
8 8 wn_synonymy
11 9 exact
12 10 exact
13 11 exact
14 12 exact

CMU + BBLT analysis on RT & MT 6 (categories i, vi, vii)
RT: the army commander found out about the start date of the battle on wednesday
MT 6: the flag of the army commander on beginning the battle on wednesday

Figure 5f. CMU analysis and CMU+BBLT analysis on on RT & MT6



علم	قائد	الجيش	بتاريخ	ابتداء	المعركة	يوم	الاربعاء
(he/it) know/find out	leader/commander	the army/troops	by/with date/history	beginning/start	the battle/campaign	today	the Wednesday
(he/it) be known/be found out	driver/chauffeur/pilot			beginning from/as of/starting with		day	
(he/it) teach/instruct	driving/leading					Youm	
knowledge/knowing							
science/study of							
flag/banner/badge							

Table 4. BBLT Results Screen for input sequence in Fig. 4, where the original input is right-to-left on the input line, but the results table reverses words into left-to-right order.

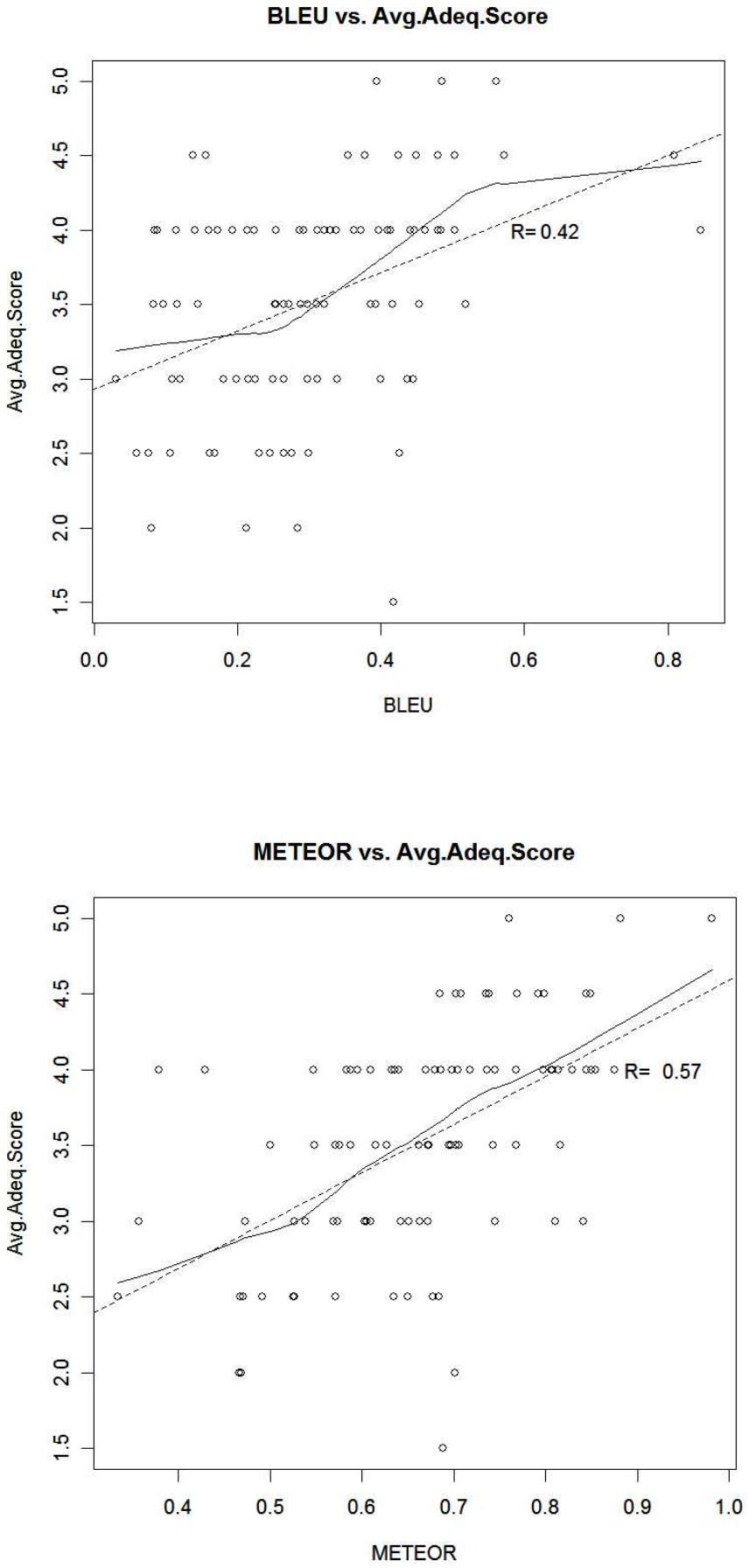


Figure 6. Distribution of Averaged Subjective Adequacy Scores By BLEU and By METEOR Scores on Output of three DARPA GALE Arabic-English MT Engines