A Human Judgment Corpus and a Metric for Arabic MT Evaluation

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit and Kemal Oflazer
Carnegie Mellon University in Qatar
{hbouamor,halshikh,behrang,ko}@cmu.edu

Abstract

We present a human judgments dataset and an adapted metric for evaluation of Arabic machine translation. Our medium-scale dataset is the first of its kind for Arabic with high annotation quality. We use the dataset to adapt the BLEU score for Arabic. Our score (AL-BLEU) provides partial credits for stem and morphological matchings of hypothesis and reference words. We evaluate BLEU, METEOR and AL-BLEU on our human judgments corpus and show that AL-BLEU has the highest correlation with human judgments. We are releasing the dataset and software to the research community.

1 Introduction

Evaluation of Machine Translation (MT) continues to be a challenging research problem. There is an ongoing effort in finding simple and scalable metrics with rich linguistic analysis. A wide range of metrics have been proposed and evaluated mostly for European target languages (Callison-Burch et al., 2011; Macháček and Bojar, 2013). These metrics are usually evaluated based on their correlation with human judgments on a set of MT output. While there has been growing interest in building systems for translating into Arabic, the evaluation of Arabic MT is still an under-studied problem. Standard MT metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006) have been widely used for evaluating Arabic MT (El Kholy and Habash, 2012). These metrics use strict word and phrase matching between the MT output and reference translations. For morphologically rich target languages such as Arabic, such criteria are too simplistic and inadequate. In this paper, we present: (a) the first human judgment dataset for Arabic MT (b) the Arabic Language BLEU (AL-BLEU), an extension of the BLEU score for Arabic MT evaluation.

Our annotated dataset is composed of the output of six MT systems with texts from a diverse set of topics. A group of ten native Arabic speakers annotated this corpus with high-levels of inter- and intra-annotator agreements. Our AL-BLEU metric uses a rich set of morphological, syntactic and lexical features to extend the evaluation beyond the exact matching. We conduct different experiments on the newly built dataset and demonstrate that AL-BLEU shows a stronger average correlation with human judgments than the BLEU and METEOR scores. Our dataset and our AL-BLEU metric provide useful testbeds for further research on Arabic MT and its evaluation.[1]

2 Related Work

Several studies on MT evaluation have pointed out the inadequacy of the standard n-gram based evaluation metrics for various languages (Callison-Burch et al., 2006). For morphologically complex languages and those without word delimiters, several studies have attempted to improve upon them and suggest more reliable metrics that correlate better with human judgments (Denoual and Lepage, 2005; Homola et al., 2009).

A common approach to the problem of morphologically complex words is to integrate some linguistic knowledge in the metric. METEOR (Denkowski and Lavie, 2011), TER-Plus (Snover et al., 2010) incorporate limited linguistic resources. Popović and Ney (2009) showed that n-gram based evaluation metrics calculated on POS sequences correlate well with human judgments, and recently designed and evaluated MPF, a BLEU-style metric based on morphemes and POS tags (Popović, 2011). In the same direc-

[1] The dataset and the software are available at: http://nlp.qatar.cmu.edu/resources/AL-BLEU

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 207–213,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

tion, Chen and Kuhn (2011) proposed AMBER, a modified version of BLEU incorporating recall, extra penalties, and light linguistic knowledge about English morphology. Liu et al. (2010) propose TESLA-M, a variant of a metric based on n-gram matching that utilizes light-weight linguistic analysis including lemmatization, POS tagging, and WordNet synonym relations. This metric was then extended to TESLA-B to model phrase synonyms by exploiting bilingual phrase tables (Dahlmeier et al., 2011). Tantug et al. (2008) presented BLEU+, a tool that implements various extension to BLEU computation to allow for a better evaluation of the translation performance for Turkish.

To the best of our knowledge the only human judgment dataset for Arabic MT is the small corpus which was used to tune parameters of the METEOR metric for Arabic (Denkowski and Lavie, 2011). Due to the shortage of Arabic human judgment dataset, studies on the performance of evaluation metrics have been constrained and limited. A relevant effort in this area is the upper-bound estimation of BLEU and METEOR scores for Arabic MT output (El Kholy and Habash, 2011). As part of its extensive functionality, the AMEANA system provides the upper-bound estimate by an exhaustive matching of morphological and lexical features between the hypothesis and the reference translations. Our use of morphological and lexical features overlaps with the AMEANA framework. However, we extend our partial matching to a supervised tuning framework for estimating the value of partial credits. Moreover, our human judgment dataset allows us to validate our framework with a large-scale gold-standard data.

3 Human judgment dataset

We describe here our procedure for compiling a diverse Arabic MT dataset and annotating it with human judgments.

3.1 Data and systems

We annotate a corpus composed of three datasets: (1) the standard English-Arabic NIST 2005 corpus, commonly used for MT evaluations and composed of news stories. We use the first English translation as the source and the single corresponding Arabic sentence as the reference. (2) the MEDAR corpus (Maegaard et al., 2010) that consists of texts related to the climate change with four Arabic reference translations. We only use the first reference in this study. (3) a small dataset of Wikipedia articles (WIKI) to extend our corpus and metric evaluation to topics beyond the commonly-used news topics. This sub-corpus consists of our in-house Arabic translations of seven English Wikipedia articles. The articles are: *Earl Francis Lloyd*, *Western Europe*, *Citizenship*, *Marcus Garvey*, *Middle Age translation*, *Acadian*, *NBA*. The English articles which do not exist in the Arabic Wikipedia were manually translated by a bilingual linguist.

Table 1 gives an overview of these sub-corpora characteristics.

	NIST	**MEDAR**	**WIKI**
# of Documents	100	4	7
# of Sentences	1056	509	327

Table 1: Statistics on the datasets.

We use six state-of-the-art English-to-Arabic MT systems. These include four research-oriented phrase-based systems with various morphological and syntactic features and different Arabic tokenization schemes and also two commercial off-the-shelf systems.

3.2 Annotation of human judgments

In order conduct a manual evaluation of the six MT systems, we formulated it as a ranking problem. We adapt the framework used in the WMT 2011 shared task for evaluating MT metrics on European language pairs (Callison-Burch et al., 2011) for Arabic MT. We gather human ranking judgments by asking ten annotators (each native speaker of Arabic with English as a second language) to assess the quality of the English-Arabic systems, by ranking sentences relative to each other, from the best to the worst (ties are allowed).

We use the Appraise toolkit (Federmann, 2012) designed for manual MT evaluation. The tool displays to the annotator, the source sentence and translations produced by various MT systems. The annotators received initial training on the tool and the task with ten sentences. They were presented with a brief guideline indicating the purpose of the task and the main criteria of MT output evaluation.

Each annotator was assigned to 22 ranking tasks. Each task included ten screens. Each screen involveed ranking translations of ten sentences. In total, we collected $22,000$ rankings for 1892 sentences (22 tasks×10 screens×10 judges). In each annotation screen, the annotator was shown the source-language (English) sentences, as well as five translations to be ranked. We did not provide annotators with the reference to avoid any bias in the annotation process. Each source sentence was presented with its direct context. Rather than attempting to get a complete ordering over the systems, we instead relied on random selection and a reasonably large sample size to make the comparisons fair (Callison-Burch et al., 2011).

An example of a source sentence and its five translations to be ranked is given in Table 2.

3.3 Annotation quality and analysis

In order to ensure the validity of any evaluation setup, a reasonable of *inter-* and *intra-*annotator agreement rates in ranking should exist. To measure these agreements, we deliberately reassigned 10% of the tasks to second annotators. Moreover, we ensured that 10% of the screens are redisplayed to the same annotator within the same task. This procedure allowed us to collect reliable quality control measure for our dataset.

	κ_{inter}	κ_{intra}
EN-AR	0.57	0.62
Average EN-EU	0.41	0.57
EN-CZ	0.40	0.54

Table 3: Inter- and intra-annotator agreement scores for our annotation compared to the average scores for five English to five European languages and also English-Czech (Callison-Burch et al., 2011).

We measured head-to-head pairwise agreement among annotators using Cohen's kappa (κ) (Cohen, 1968), defined as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times annotators agree and P(E) is the proportion of agreement by chance.

Table 3 gives average values obtained for inter-annotator and intra-annotator agreement and compare our results to similar annotation efforts in WMT-13 on different European languages. Here we compare against the average agreement for English to five languages and also from English to one morphologically rich language (Czech).[4]

Based on Landis and Koch (1977) κ interpretation, the κ_{inter} value (57%) and also comparing our agreement scores with WMT-13 annotations, we believe that we have reached a reliable and consistent annotation quality.

4 AL-BLEU

Despite its well-known shortcomings (Callison-Burch et al., 2006), BLEU continues to be the de-facto MT evaluation metric. BLEU uses an exact n-gram matching criterion that is too strict for a morphologically rich language like Arabic. The system outputs in Table 2 are examples of how BLEU heavily penalizes Arabic. Based on BLEU, the best hypothesis is from Sys_5 which has three unigram and one bigram exact matches with the reference. However, the sentence is the 4^{th} ranked by annotators. In contrast, the output of Sys_3 (ranked 1^{st} by annotators) has only one exact match, but several partial matches when morphological and lexical information are taken into consideration.

We propose the Arabic Language BLEU (AL-BLEU) metric which extends BLEU to deal with Arabic rich morphology. We extend the matching to morphological, syntactic and lexical levels with an optimized partial credit. AL-BLEU starts with the exact matching of hypothesis tokens against the reference tokens. Furthermore, it considers the following: (a) morphological and syntactic feature matching, (b) stem matching. Based on Arabic linguistic intuition, we check the matching of a subset of 5 morphological features: (i) POS tag, (ii) gender (iii) number (iv) person (v) definiteness. We use the MADA package (Habash et al., 2009) to collect the stem and the morphological features of the hypothesis and reference translation.

Figure 1 summarizes the function in which we consider partial matching ($m(t_h, t_r)$) of a hypothesis token (t_h) and its associated reference token (t_r). Starting with the BLEU criterion, we first check if the hypothesis token is same as the reference one and provide the full credit for it. If the exact matching fails, we provide partial credit for matching at the stem and morphological level. The value of the partial credits are the sum of the stem weight (w_s) and the morphological fea-

[4] We compare against the agreement score for annotations performed by WMT researchers which are higher than the WMT annotations on Mechanical Turk.

Source	France plans to attend ASEAN emergency summit.						
Reference	فرنسا تعتزم حضور قمة الاسيان الطارئة. *frnsaA tEtzm HDwr qmp AaAlaAsyaAn AaAlTaAr}ip*						
Hypothesis	**Systems**	**Rank**$_{Annot}$	**BLEU**	**Rank**$_{BLEU}$	**AL-BLEU**	**Rank**$_{AL-BLEU}$	
	Sys$_1$	2	0.0047	2	**0.4816**	**1**	وتخطط فرنسا لحضور قمة الآسيان الطارئة *wtxTaT frnsaA lHDwr qmp AaAl—syaAn AaAlTaAr}ip*
	Sys$_2$	3	0.0037	3	0.0840	3	وتخطط فرنسا لحضور قمة الأسيان *wtxTaT frnsaA lHDwr qmp AaAlOasyaAn*
	Sys$_3$	1	0.0043	4	0.0940	2	فرنسا تخطط لحضور القمة الطارئة للأسيان *frnsaA txTaT lHDwr AaAlqmp AaAlTaAr}ip lalOasyaAn*
	Sys$_4$	5	0.0043	4	0.0604	5	خطط فرنسا لحضور قمة آسيان الطوارئ *xTaT frnsaA lHDwr qmp —syaAn AaAlTwaAri}*
	Sys$_5$	4	**0.0178**	**1**	0.0826	4	فرنسا لحضور قمة الاسيان خطط الطوارئ *frnsaA lHDwr qmp AaAlaAsyaAn xTaT AaAlTwaAri}*

Table 2: Example of ranked MT outputs in our gold-standard dataset. The first two rows specify the English input and the Arabic reference, respectively. The third row of the table lists the different MT system as ranked by annotators, using BLEU scores (column 4) and AL-BLEU (column 6). The different translation candidates are given here along with their associated Bucklwalter transliteration.[3] This example, shows clearly that AL-BLEU correlates better with human decision.

$$m(t_h, t_r) = \begin{cases} 1, & \textbf{if } t_h = t_r \\ w_s + \sum_{i=1}^{5} w_{fi} & \textbf{otherwise} \end{cases}$$

Figure 1: Formulation of our partial matching.

ture weights (w_{fi}). Each weight is included in the partial score, if such matching exist (e.g., stem match). In order to avoid over-crediting, we limit the range of weights with a set of constraints. Moreover, we use a development set to optimize the weights towards improvement of correlation with human judgments, using a hill-climbing algorithm (Russell and Norvig, 2009). Figure 2 illustrates these various samples of partial matching highlighted in different colors.

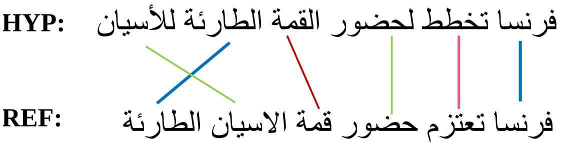
HYP: فرنسا تخطط لحضور القمة الطارئة للأسيان
REF: فرنسا تعتزم حضور قمة الاسيان الطارئة

Figure 2: An MT example with exact matchings (blue), stem and morphological matching (green), stem only matching (red) and morphological-only matching (pink).

Following the BLEU-style exact matching and scoring of different n-grams, AL-BLEU updates the n-gram scores with the partial credits from non-exact matches. We use a minimum partial credit for n-grams which have tokens with different matching score. The contribution of a partially-matched n-gram is not 1 (as counted in BLEU), but the minimum value that individual tokens within the bigram are credited. For example, if a bigram is composed of a token with exact matching and a token with stem matching, this bigram receives a credit equal to a unigram with the stem matching (a value less than 1). While partial credits are added for various n-grams, the final computation of the AL-BLEU is similar to the original BLEU based on the geometric mean of the different matched n-grams. We follow BLEU in using a very small smoothing value to avoid zero n-gram counts and zero score.

5 Experiments and results

An automatic evaluation metric is said to be successful if it is shown to have high agreement with human-performed evaluations (Soricut and Brill, 2004). We use Kendall's tau τ (Kendall, 1938), a coefficient to measure the correlation between the system rankings and the human judgments at the sentence level. Kendall's tau τ is calculated as follows:

$$\tau = \frac{\text{\# of concordant pairs - \# of discordant pairs}}{\text{total pairs}}$$

where a *concordant pair* indicates two translations of the same sentence for which the ranks obtained from the manual ranking task and from the corresponding metric scores agree (they disagree in a *discordant pair*). The possible values of τ range from -1 (all pairs are discordant) to 1 (all pairs

	Dev	**Test**
BLEU	0.3361	0.3162
METEOR	0.3331	0.3426
AL-BLEU$_{Morph}$	**0.3746**	0.3535
AL-BLEU$_{Lex}$	0.3732	**0.3564**
AL-BLEU	**0.3759**	**0.3521**

Table 4: Comparison of the average Kendall's τ correlation.

are concordant). Thus, an automatic evaluation metric with a higher τ value is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower τ. We calculate the τ score for each sentence and average the scores to reach the corpus-level correlation. We conducted a set of experiments to compare the correlation of AL-BLEU against the state-of-the art MT evaluation metrics. For this we use a subset of 900 sentences extracted from the dataset described in Section 3.1. As mentioned above, the stem and morphological features in AL-BLEU are parameterized each by weights which are used to calculate the partial credits. We optimize the value of each weight towards correlation with human judgment by hill climbing with 100 random restarts using a development set of 600 sentences. The 300 remaining sentences (100 from each corpus) are kept for testing. The development and test sets are composed of equal portions of sentences from the three sub-corpora (NIST, MEDAR, WIKI).

As baselines, we measured the correlation of BLEU and METEOR with human judgments collected for each sentence. We did not observe a strong correlation with the Arabic-tuned METEOR. We conducted our experiments on the standard METEOR which was a stronger baseline than its Arabic version. In order to avoid the zero n-gram counts and artificially low BLEU scores, we use a smoothed version of BLEU. We follow Liu and Gildea (2005) to add a small value to both the matched n-grams and the total number of n-grams (epsilon value of 10^{-3}). In order to reach an optimal ordering of partial matches, we conducted a set of experiments in which we compared different orders between the morphological and lexical matchings to settle with the final order which was presented in Figure 1.

Table 4 shows a comparison of the average correlation with human judgments for BLEU, METEOR and AL-BLEU. AL-BLEU shows a strong improvement against BLEU and a competitive improvement against METEOR both on the test and development sets. The example in Table 2 shows a sample case of such improvement. In the example, the sentence ranked the highest by the annotator has only two exact matching with the reference translation (which results in a low BLEU score). The stem and morphological matching of AL-BLEU, gives a score and ranking much closer to human judgments.

6 Conclusion

We presented AL-BLEU, our adaptation of BLEU for the evaluation of machine translation into Arabic. The metric uses morphological, syntactic and lexical matching to go beyond exact token matching. We also presented our annotated corpus of human ranking judgments for evaluation of Arabic MT. The size and diversity of the topics in the corpus, along with its relatively high annotation quality (measured by IAA scores) makes it a useful resource for future research on Arabic MT. Moreover, the strong performance of our AL-BLEU metric is a positive indicator for future exploration of richer linguistic information in evaluation of Arabic MT.

7 Acknowledgements

We thank Michael Denkowski, Ahmed El Kholy, Francisco Guzman, Nizar Habash, Alon Lavie, Austin Matthews, Preslav Nakov for their comments and help in creation of our dataset. We also thank our team of annotators from CMU-Qatar. This publication was made possible by grants YSREP-1-018-1-004 and NPRP-09-1140-1-177 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In

Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland.

Boxing Chen and Roland Kuhn. 2011. AMBER: A Modified BLEU, Enhanced Ranking Metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 71–77, Edinburgh, Scotland.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. TESLA at WMT 2011: Translation Evaluation and Tunable Metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 78–84, Edinburgh, Scotland, July. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, UK.

Etienne Denoual and Yves Lepage. 2005. BLEU in Characters: Towards Automatic MT Evaluation in Languages Without Word Delimiters. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, Jeju Island, Republic of Korea.

Ahmed El Kholy and Nizar Habash. 2011. Automatic Error Analysis for Morphologically Rich Languages. In *Proceedings of the MT Summit XIII*, pages 225–232, Xiamen, China.

Ahmed El Kholy and Nizar Habash. 2012. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. *Machine Translation*, 26(1):25–45.

Christian Federmann. 2012. Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98(1):25–35.

N. Habash, O. Rambow, and R. Roth. 2009. Mada+Tokan: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

Petr Homola, Vladislav Kuboň, and Pavel Pecina. 2009. A Simple Automatic MT Evaluation Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 33–36, Athens, Greece, March. Association for Computational Linguistics.

Maurice G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation Evaluation of Sentences with Linear-Programming-Based Analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 354–359.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria.

Bente Maegaard, Mohamed Attia, Khalid Choukri, Olivier Hamon, Steven Krauwer, and Mustafa Yaseen. 2010. Cooperation for Arabic Language Resources and Tools–The MEDAR Project. In *Proceedings of LREC*, Valetta, Malta.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Maja Popović and Hermann Ney. 2009. Syntax-oriented Evaluation Measures for Machine Translation Output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32, Athens, Greece.

Maja Popović. 2011. Morphemes and POS Tags for n-gram Based Evaluation Metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 104–107, Edinburgh, Scotland.

Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach.* Prentice Hall Englewood Cliffs.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, Boston, USA.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2010. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

Radu Soricut and Eric Brill. 2004. A Unified Framework For Automatic Evaluation Using 4-Gram Co-occurrence Statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 613–620, Barcelona, Spain, July.

Cüneyd Tantug, Kemal Oflazer, and Ilknur Durgar El-Kahlout. 2008. BLEU+: a Tool for Fine-Grained BLEU Computation. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco.