Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study

Laith S. Hadla
Department of Translation, Faculty of Arts, Zarqa University
Zarqa - Jordan

Taghreed M. Hailat
Faculty of IT and CS
Yarmouk University
Irbid, Jordan

Mohammed N. Al-Kabi
Computer Science Department
Zarqa University
P.O. Box 132222
Zarqa 13132, Jordan

***Abstract*—The Internet provides its users with a variety of services, and these services include free online machine translators, which translate free of charge between many of the world's languages such as Arabic, English, Chinese, German, Spanish, French, Russian, etc. Machine translators facilitate the transfer of information between different languages, thus eliminating the language barrier, since the amount of information and knowledge available varies from one language to another, Arabic content on the internet, for example, accounts 1% of the total internet content, while Arabs constitute 5% of the population of the earth, which means that the intellectual productivity of the Arabs is low because within internet use Internet's Arabic content represents 20% of their natural proportion, which in turn encouraged some Arab parties to improve Arabic content within the internet. So, many of those interested specialists rely on machine translators to bridge the knowledge gap between the information available in the Arabic language and those in other living languages such as English.**

This empirical study aims to identify the best Arabic to English Machine translation system, in order to help the developers of these systems to enhance the effectiveness of these systems. Furthermore, such studies help the users to choose the best. This study involves the construction of a system for Automatic Machine Translation Evaluation System of the Arabic language into language. This study includes assessing the accuracy of the translation by the two known machine translators, Google Translate, and the second, which bears the name of Babylon machine translation from Arabic into English. BLEU and METEOR methods are used the MT quality, and to identify the closer method to human judgments. The authors conclude that BLEU is closer to human judgments METEOR method.

Keywords—component; Machine Translation; Arabic-English Corpus; Google Translator; Babylon Translator; METEOR; BLEU

I. INTRODUCTION

The term Machine Translation (MT) dates back to the 1950's., and it is one of the earliest areas of research within Natural Language Processing (NLP) field. Until this moment, the accuracy of machine translation is lower than that of professional translators. There are different methods to translate from one natural language into another, and these methods are adopted by Online Machine Translation Systems such as Statistical Machine Translation (SMT), Hybrid Machine Translation (HMT), Rule-based, Knowledge-based, Interlingua, Direct, Transfer, and Example-based. The first two approaches (SMT and HMT) are the most widely used approaches nowadays.

The professional human translators are the best to evaluate the translation quality of Machine Translation Systems, but this way costs time, money, and effort consuming as the human translation. Therefore, many new methods are proposed by researchers to automatically evaluate the quality of the output of Machine Translation Systems. The utilization of these methods is not constrained to the automatic evaluation of MT systems, but it can be used in Software Development Life Cycle (SDLC) of MT systems, to enhance the efficiency of software under construction, analyze errors, and MT system benchmark. All these automatic MT evaluation methods depend on a core idea of making a comparison of the corresponding candidate translations and reference translations. We have to consider the fact that the correct human translation is not unique, and the list of valid reference translations is not limited. Therefore, this type of evaluation is considered a subjective, since it is highly correlated to human judgments (reference translations), and this leads to the difficulty. Manual (human) evaluation of MT is characterized by direct interpretation and accuracy relative to automatic evaluation of MT, but it costs money and time relative to automatic evaluation of MT.

Furthermore, the disadvantages of manual evaluation are non-reusability and subjectivity. On the other hand, automatic evaluation of MT is characterized by reusability, speed and free of charge, and it has a list of cons presented in the literature. The first generation of automatic MT evaluation methods depends on lexical similarity (n-gram -based) measures to compute their scores that represent the lexical matching between corresponding candidate Translations and reference translations [1].

Two widely used methods to automatically evaluate MT systems are used in this study. BLEU (Bilingual Evaluation Understudy) method is one of the earlier methods cast in this field, and it is used in this study. Therefore, as noted before, that the earlier methods of automatic MT evaluation depend on lexical similarity (n-gram -based) measures to compute their scores. BLEU score value is between 0 and 1, where 1

This research is funded by the Deanship of Research at Zarqa University / Jordan.

indicates the candidate translation and reference translation are fully matched, and 0 indicates the candidate translation and reference translation are completely different. BLEU values close to 1 indicates the similarity of the two translations is high, and BLEU values close to 0 indicate the similarity of the two translations is low. Those who use BLEU can benefit from its language independence and high correlation with human judgment. Furthermore, as other similar methods it has its pitfalls [1], [2]. The core idea of this widely used method based on the use of modified ngram precision, and so it needs to compute the number of common n-grams in the corresponding candidate and reference translations regardless of the position of matched n-grams. Then, the number of common words is divided by the total number of words in the candidate translation.

The second method used in this study is called METEOR 1.5 (Metric for Evaluation of Translation with Explicit ORdering) is an automatic evaluation metric for the machine translation output. Lavie, Kenji and Jayaraman study [3] proposes and casts METEOR metric for the first time in 2004, and aimed to improve correlation with human judgments of MT quality at the segment level. METEOR scores machine translation hypotheses by aligning them to one or more reference translations. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases. METEOR has several features that are not found in BLEU, such as stemming and synonym matching, along with the standard exact word matching. On the other hand, BLEU and NIST metrics are based on precision alone, but the METEOR metric mentioned before uses precision and recall. Researchers proved that using precision and recall by METEOR leads to a higher correlation with human judgment at the sentence or segment level relative to metrics like BLEU and NIST [4]. Furthermore, METEOR score includes a fragmentation penalty that considers how well ordered the matched unigrams of the candidate translation are with respect to the reference. Carnegie Mellon University releases five versions of METEOR (Version 1.0, Version 1.2, Version 1.3, Version 1.4, and Version 1.5) on its Web page (http://www.cs.cmu.edu/~alavie/METEOR/).

Semitic languages include the following list of languages sorted according to native speakers: Arabic, Amharic, Hebrew, Aramaic, etc. The number of Arabic native speakers is widely varied from 220 to 400million people, besides Muslims who use it during the practice of their faith [5]. The Arabic language is one of the official languages used by all member states of the Arab league, and it is one of the UN official languages. Nowadays, two types of Arabic language are used, the first type is called Modern Standard Arabic (MSA) and it used mainly in official correspondence, books, journals, newspapers, etc., the second type includes a number of varying Arabic dialects that used in homes, markets, cafes, chatting, etc. Therefore, the spoken vernaculars are varied from country to country, and sometimes from village to a near village. The Arabic language is different from the English language since it has 28 Arabic letters, written from right to left in cursive style. The shape of the letter inside a word depends on its position (initial, medial, or final). The Arabic language lacks to the capitalization found in the English language [6].

This research is funded by the Deanship of Research at Zarqa University / Jordan.

Many studies exhibit different methods to improve MT of Arabic into other languages [7], [8], [9], [10], and [11].

In this study, researchers set out to conduct experiments that were used to benchmark the two methods (BLEU and METEOR 1.5) used for MT evaluation, and the two free online MT systems (Google Translate (https://translate.google.com) and Babylon system (http://translation.babylon.com/)) used to translate the 1033 Arabic sentences into English. Therefore, this study includes building an Automatic Machine Translation Evaluation System from Arabic into English using METEOR 1.5 and BLEU methods. The data set is divided to the sentence types (imperative, declarative, exclamatory, and interrogative). It is to be mentioned that these sentences were used in a previous study [12].

This paper is structured as follows: Section II reviews the related studies on automatic MT evaluation, and specifically those related to METEOR and BLEU methods. In Section III, we describe the framework and methodology of this study. In Section IV, we present our experimental results of the system designed and implemented by the second author and the results of two free online machine translation systems using a small data set consisting of 1033 Arabic sentences. Section V presents the conclusion from this paper. Finally, section VI presents plans to extend and improve this paper.

II. Related Work

Kirchhoff, Capurro, and Turner study [13] categorize the evaluation of machine translation (MT) into three main categories: human evaluation category, automatic evaluation category, and embedded application evaluation category.

This section starts with presenting studies related to Metric for Evaluation of Translation with Explicit Ordering (METEOR) method to automatically evaluate machine translation. Afterward, papers related to (BLEU) method are presented. Last and not least this section presents eleven studies related to the automatic evaluation of MT that includes Arabic.

Lavie, Kenji and Jayaraman in their study [3] cast a new metric in 2004 called METEOR to automatically evaluate MT systems. Some of the deficiencies of the BLEU score attempted to be addressed by METEOR metric. METEOR is based on a generalized concept of unigram matching between the Candidate Translation and Reference translation. METEOR flexibly matches unigrams using stemming and WordNet synonyms that does not require exact matching of words between the Candidate Translation and the Reference translation. This first metric METEOR attempts to determine all generalized unigram matches between the candidate translation and reference translation, then it starts computing METEOR score using a combination of unigram-precision, unigram-recall, and a measure of fragmentation to measure how well-ordered the matched words in the candidate translation are in relation to the reference. Banerjee and Lavie study [4] tests METEOR using LDC TIDES 2003 Arabic-into-English and Chinese-into-English data sets, to prove it yields better results than its counterparts (BLEU, NIST, Precision, Recall, F_1, and F_{mean}). In other words, they prove that

METEOR is closer to human judgments relative other metrics using the Arabic and Chinese data sets. Lavie and Agarwal study [14] presents an improvement to METEOR presented in [3] and [4]. This improved version supports additional languages like Spanish, French, and German, in addition to English, Arabic, and Chinese [14]. METEOR metric accounts for reordering to enhance the correlation with human judgments of translation accuracy [15].

Denkowski and Lavie [16] present in their study an improvement to METEOR metric. The new improved version is called METEOR-NEXT, and it includes METEOR's original features, besides paraphrases; and more sophisticated metrics that use deeper linguistic information.

Another study by Denkowski and Lavie [17] presents an improvement to METEOR metric. The new METEOR 1.3 includes improved text normalization, higher-precision paraphrase matching, and discrimination between content and function words. Furthermore, this improved version of METEOR metric includes ranking and adequacy versions and overcome some weaknesses of previous versions of METEOR such as noise in the paraphrase matching, lack of punctuation handling and discrimination between word types.

The second part of this section presents studies related to (BLEU) as a second method used to automatically evaluate MT quality.

It is mentioned before in this study that the number of valid reference translations of a certain source text is not limited to one, two, three, etc. valid reference translations. Based on this fact in 2002 Papineni et al. [2] cast BLEU to automatically evaluate the accuracy of the output of MT system using one, two, or more reference translations beside the corresponding candidate translation. BLEU is an n-gram based metric, where n ranges from 1 to 4. BLEU scores are highly affected by the number of reference translations, and that means the more reference translations per candidate translation there are, the higher BLEU score is. Therefore, BLEU requires multiple good reference translations.

Modified n-gram precision is an improved version of n-gram precision that aims to identify and avoid rewarding false positives outputted by MT systems. Brevity penalty is a correcting factor used to prevent short candidate translations relative to their reference counterparts from receiving a high a score. Small variations in candidate and reference translation lengths have a small impact on overall BLEU score. BLEU score is a product of multiplying modified n-gram precision at a sentence level by brevity penalty factor.

Although, many previous studies propose an enhanced BLEU method, only three studies are presented in this section due to space limitation.

Babych and Hartley [18] presents a modified version of BLEU method which uses Weighted n-gram Model that depends on a modified version of tf -idf to compute the weights of different words according to their importance, and S-score weights are used for translation Adequacy and Fluency. The S-score helps to weigh Content words differently from common words. DARPA-94 MT French-English evaluation corpus, which has 100 news text is used in their study to evaluate the effectiveness of the enhanced BLEU model. In their experiments, they used five MT systems to translate each of the 100 French news texts into English, and four of these candidate translations are evaluated by professional human translators. The corpus used by them has 2 English reference translations for each of the 100 French news texts. The results of their experiments reveal that their enhanced method scores for fluency are consistent with the base-line BLEU scores for fluency, but their enhanced method scores for adequacy outperform the base-line BLEU scores for adequacy. This modified version of BLEU can use only one reference translation, and yields a reliable result.

Another proposed extended BLEU method is presented in a study conducted by Yang, Zhu, Li, Wang, Qi, Li and Daxin [19]. They proposed assigning different weights to different part-of-speech (POS) and different lengths of n-gram. The information related to POS and lengths of n-gram are introduced to a linear regression model within the classical BLEU framework. This extension to BLEU does not affect the language independence of the original BLEU. Experimental results of the extended BLEU method show it is more effective than the baseline BLEU method.

An extended version of BLEU called AMBER is presented in Chen and Kuhn [20] study. The extended version of BLEU includes several new penalties instead of the brevity penalty used in the original BLEU. Furthermore, the computation of their metric includes text processing operations and the use of F-measure instead of precision and, therefore, they have to compute recall and precision before computing the F-measure. AMBER test results show it is more effective than the original BLEU metric. There is relatively little number of studies in the literature concerned with the evaluation of Arabic MT systems.

Therefore, Guessoum and Zantout [21] decided to evaluate four English-Arabic commercial MT systems (ATA, Arabtrans, Ajeeb, and Al-Nakel) using their new proposed to evaluate MT systems. The evaluation results show poor performance generally, except the lexical coverage of the domain of the Internet and Arabization.

Al-Haj and Lavie [22] study refers to the challenges facing statistical machine translation (SMT) such as Google Translate to translate from or into a morphologically rich language, and this challenge is magnified when translating into a morphologically rich language like Arabic. They addressed this challenge in the framework of a detailed description English-to-Arabic phrase-based statistical machine translation (PBSMT). Morphological segmentation and tokenization decisions have a great impact of the effectiveness of English-to-Arabic PBSMT outputs. Al-Haj and Lavie [22] present BLEU scores of different morphological segmentation schemes. Therefore, they deduce that a proper choice of segmentation has a significant effect on the performance of the SMT.

All the studies that use BLEU and METEOR methods use reference translations, in addition to candidate translations, but an interesting study conducted by Palmer [23] to automatically evaluate candidate translations depends on user-centered method and do not reply on reference translations. Palmer's method compares the outputs of MT systems and then ranking

them, according to their quality. A number of professional users who have the necessary linguistics skills are used to rank candidate translations. The tests of Palmer's method [23] include seven MT systems, four Arabic-into-English MT systems, and three Mandarin (simplified Chinese)-into-English MT systems. Palmer study [23] is based on spoken language transcripts, and not on a textual data set.

Arabic dialects are the real languages used by most of the people in the Arab world to communicate with each other at homes, markets, restaurants, hospitals, etc. Arabs use many dialects that vary from a place to another, and Iraqi Arabic is one of these dialects that used in Iraq as the name indicates, and it is close to dialects used in the gulf region. Condon et al. study [24] presents an automatic method to evaluate the quality of Iraqi Arabic-English speech translation dialogues. They show that normalization has a positive effect on making the candidate translations closer to human judgments.

Adly and Al-Ansary [10] study presents an evaluation of an MT system that based on the interlingua approach, and Multilanguage MT system called Universal Network Language (UNL) system.

They address in their study [10] the evaluation of English-Arabic MT using three metrics BLEU, F_1 and F_{mean}, and conclude that UNL MT accuracy outperforms other MT systems. Alansary, Nagi, and Adly [25], and Al-Ansary [26] studied the effect of UNL MT system on translation from/into the Arabic language.

Carpuat, Marton, and Habash's [7] study overlaps with our study since it is concerned with translation from Arabic into English. They addressed in their study three main challenges: reordering, subject detection, and Arabic verb in Statistical Machine Translation. Furthermore, Carpuat, Marton, and Habash's [7] proposed a reordering of Verb Subject (VS) construction into Subject Verb (SV) construction for alignment only to minimize ambiguities. The results of their proposal show an improvement in BLEU and TER scores.

Alqudsi, Omar, and Shaker [27] conducted a good survey about available MT techniques, and exhibited some of the linguistic characteristics of the Arabic language with an emphasis on linguistic characteristics that have negative effects on MT. The study [27] presents a summary of the strengths and weaknesses of the main methods used in MT from Arabic into English.

A preliminary study is conducted by Hailat, Al-Kabi, Alsmadi, and Shawakfa [28] to evaluate the effectiveness of translation from English into Arabic of two free online MT systems (Google Translate and Babylon machine translation systems). They used a small data set that consists of 200 English sentences. BLEU was used to automatically evaluate the accuracy of each system. The evaluation results indicate Google Translate system is more effective than its counterpart.

The authors of the previous study [28] decide to improve their study using a larger data set of English sentences relative to the data set used in their previous study. They used in their new study [29] the same two online MT systems (Google Translate (https://translate.google.com) & Babylon (http://translation.babylon.com/) they used before, and they use the same method (BLEU) used before to automatically evaluate MT systems. They conclude that Google Translate is generally more accurate than its counterpart.

ATEC is another metric usually used to automatically evaluate the outputs of MT systems. The effectiveness of each of the 2 free Online Machine Translation systems "IMTranslator" and "Google Translate MT system" is explored by Al-Deek, Al-Sukhni, Al-Kabi, and Haidar [30] to conclude the Google Translate is more accurate than its counterpart.

A closely related study to our is conducted by Hadla, Hailat, and Al-Kabi study [12], to identify the best of two online machine translation systems (Google Translate and Babylon MT systems) to translate from Arabic into English. BLEU method is used by these authors to evaluate translation effectiveness of the above two online MT systems under consideration. They used more than 1000 Arabic sentences in their study to conduct their benchmark, where each Arabic sentence is accompanied by 2 reference English translations. The Arabic sentences they used are classified into four classes, where each class represents one of the four basic sentence functions (declarative, exclamatory, interrogative, and imperative). Hadla, Hailat, and Al-Kabi study [12] study concludes that Google Translate system is more accurate than Babylon MT system in terms of translation from Arabic into English.

III. The Methodology

This study is based on a data set constructed by Hadla, Hailat, and Al-Kabi [12], that consists of 1033 Arabic sentences with two reference translations of each Arabic sentence in the data set. This is an open access data set that can be downloaded from the following URL:https://docs.google.com/spreadsheets/d/1bqknBcdQ7cXOKtYLhVP7YHbvrlyJlsQggL60pnLpZfA/edit?usp=sharing

The 1033 Arabic sentences of the above data set is distributed among four basic sentence functions (declarative: 250 Arabic sentences, interrogative: 281 Arabic sentences, exclamatory: 252 Arabic sentences, and imperative: 250 Arabic sentences).

Figure 1 shows the main steps of the methodology, and how to extract n-grams from the Arabic, English Candidate, English Reference sentences, to be used to compute METEOR 1.5 and BLEU scores for Google Translate and Babylon machine translation system. Afterward, the closest score to the human judgment is determined.

METEOR method is used to automatically evaluate machine translation systems, and it uses word matching in target and reference translations to evaluate the accuracy of the machine translation. METEOR score is based mainly on word-to-word matches between target and reference translations. When more than one reference translation is available, the METEOR score is computed independently for each reference translation and the best METEOR score is adopted. METEOR consists of two main components, the first is a flexible monolingual word aligner component, and the second component is a scorer [32]. METEOR creates a word alignment between the two target and reference translations in the comparison process. Word alignment means the mapping

between words in candidate and reference translations so that every word in each translation maps to at most one word in the other translation. Word-mapping modules are used to produce incremental alignment. These modules include modules such as the "exact" module that maps two words if they are fully matched. METEOR second module called "porter stem", and this module uses Porter Stemmer to yield stems that are mapped if they are fully matched. METEOR third module called "WN synonymy", and this module uses "synset" in WordNet to yield synonyms that mapped if they are fully matched [33]. The "porter stem" and "WN synonymy" modules do not support the Arabic language; therefore, Arabic is partially supported by METEOR 1.5 word-mapping modules [32].

The BLEU-score formula is a product of multiplying Brevity Penalty (BP) by geometric average of the modified n-gram precisions. Therefore, we have to start computing the geometric average of the modified n-gram precisions. Afterward, the length of the candidate translation (c), and the length of the effective reference corpus (r) have to be computed, in order to be able to compute the Brevity Penalty (BP). Formula 1 [2] shows how to make Brevity Penalty Reduced exponentially in (r / c).

$$BP = \begin{cases} 1 & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c \le r \end{cases} \quad (1)$$

Formula 2 exhibits how to compute the final BLEU score.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (2)$$

Where $N = 4$ and uniform weights $w_n = (1/N)$ [2].

METEOR score is computed for each corresponding candidate and reference translations, and it proposed by (Banerjee, and Lavie, 2005) [4] to automatically evaluate MT. The range of METEOR score is between 0 and 1. Unlike base-line BLEU score that depends on precision only, METEOR's score uses recall in addition to the precision, with more emphasize on Recall. Furthermore, METEOR incorporates stemming and if English is the target WordNet is used to yield English synonyms.

The computation of METEOR score needs computing unigram precision (P), and unigram recall (R) first in order to be able to compute F-mean as shown in the following formula (3) [4]:

$$Fmean = \frac{10PR}{R+9P} \quad \text{.....(3)}$$

Afterward, METEOR method computes a penalty for a given alignment as shown in the following formula (4) [4]:

$$Penalty = 0.5\left(\frac{\#chunks}{\#unigrams\_matched}\right) \quad \text{.....(4)}$$

The formula of computing the final METEOR score is shown in the following formula (5) [4]:

$$METEOR\,Score = Fmean\left(1 - Penalty\right) \quad \text{.....(5)}$$

The higher score whether it represents BLEU or METEOR means that the candidate is closer to reference translation. Therefore, the higher BLEU or METEOR score means it is closer to human judgment. METEOR assigns a score in the range of 0 to 1 to every candidate translation [31].

The values of the BLEU metric range from 0 to 1 [2]; where the value of 1 means that the candidate translation fully matched reference translation, and the value of 0 means that the candidate translation and the corresponding reference translation are completely different.

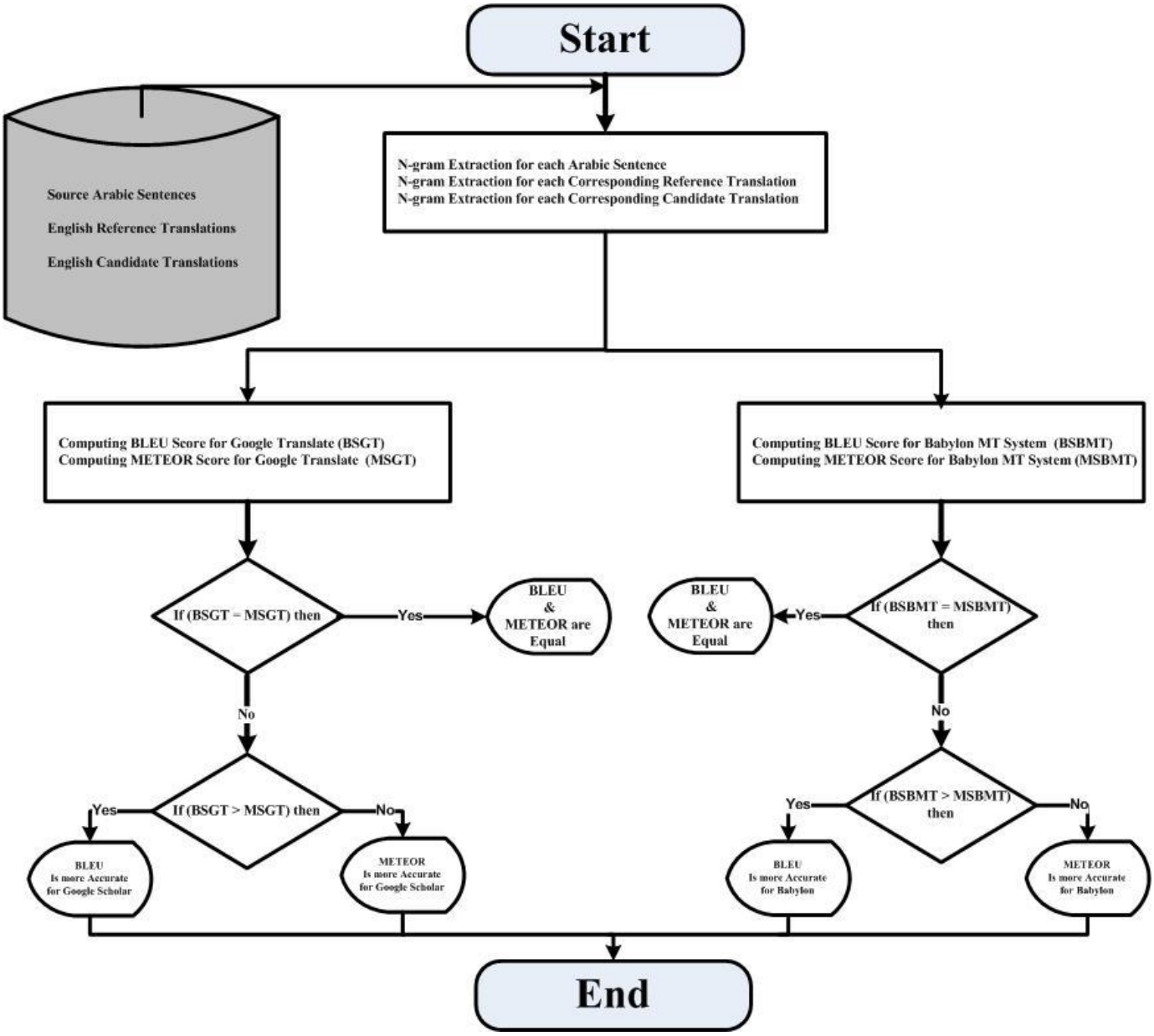

Fig. 1. Flowchart of BLEU and METEOR Evaluation Methodology

IV. The Evaluation

This study uses two well-known automatic evaluation methods (METEOR & BLEU). METEOR method measures precision and recall of unigrams when comparing a hypothesis translation against a reference one, and is characterized by its flexibility to match words (semantic matching) using stemming, paraphrase and WordNet synonyms [4].

When conducting tests on Google Translate and Babylon MT systems, the following was noted:

1) The conducted tests reveal within the outputs of Babylon MT system some Arabic words that indicate the incapability of Babylon MT system to translate these words into English, but, further investigation that these Arabic words with other inputted Arabic sentences are translated correctly into English.

Babylon MT system could not translate the words that contain related pronouns "الضمائر المتصله", for example in case of the source sentence in Arabic (I heard a word that made me laugh) " سمعتكلمةاضحكتني", was translated by Babylon MT as: "I heard the word اضحكتنى" and this note is already mentioned by the authors of [12].

TABLE I. BLEU AVERAGE PRECISION FOR EACH TYPE OF SENTENCES

Type / Translator	Declarative Sentence	Exclamation Sentence	Imperative Sentence	Interrogative Sentence	Average
Babylon MT System	0.3475	0.3686	0.5189	0.3588	0.39845
Google Translate System	0.4486	0.3378	0.5453	0.4668	0.449625

TABLE II. METEOR 1.5AVERAGE PRECISION FOR EACH TYPE OF SENTENCES

Respect to / Type	Ref1 + Google	Ref2 + Google	Ref1 + Babylon	Ref2 + Babylon
Declarative Sentence	0.39	0.32	0.33	0.30
Exclamation Sentence	0.36	0.28	0.20	0.19
Imperative Sentence	0.56	0.41	0.46	0.36
Interrogative Sentence	0.35	0.30	0.38	0.34
Average	0.415	0.328	0.343	0.298

2) Babylon machine translation system could not translate multiple Arabic sentences at one time while Google Translate has the feature of translating a set of Arabic sentences at one time.

The use of BLEU to evaluate and test these two online MT systems reveals that for some sentences the BLEU scores are equal to Google Translate & Babylon MT systems. However, the effectiveness of Google Translate system is generally better than the effectiveness of its counterpart as shown in Table 1.

Table 1 shows the evaluations of BLUE scores for each type of the 1033 Arabic sentences (imperative, declarative, exclamatory, and interrogative), and BLEU average precision values are presented for Babylon MT System and Google Translate System.

BLEU average precision values are used to identify the best MT system. Table 1 shows that Google Translate system is generally better than its counterpart, since it has a higher BLEU average precision. Furthermore, Table 1 shows Babylon MT System is better than Google Translate system in translating Arabic exclamation sentences into English. We have to note that the values of Table 1 are fully matched with those presented by [12] since we use their data set and their method (BLEU).

Table 2 shows the METEOR 1.5 average precision for the 4 types of sentences. Overall, the translations precision is below 50% except in imperative sentences that are translated by Google MT system (56%).

The values of BLEU and METEOR 1.5 methods in Table 1 and Table 2 generally show that Google Translate is more accurate than Babylon MT. Furthermore, these values do not imply that METEOR 1.5 method is more accurate than BLEU method. This deterioration in the scores of METEOR 1.5 relative to BLEU scores due to the Arabic language is not fully supported by METEOR 1.5, and therefore stems and synonyms are not used by METEOR 1.5 system.

V. CONCLUSION

In this study, two well-known Automatic MT Evaluation methods (METEOR 1.5 & BLEU) are used to identify the evaluation method that is closer to human judgment. Furthermore, this study includes tests to the effectiveness of two online MT systems (Google Translate & Babylon MT) systems to translate the 1033 Arabic sentences into English.

Most of the methods used to automatically evaluate the accuracy of the translation of MT system are based on a comparison between candidate and reference translations. This type of studies need a standard corpus, but, unfortunately, no standard corpus accepted by all researchers was found, except for an Arabic-English data set constructed and released by Hadla, Hailat, and Al-Kabi [12] to be used by the researchers was found. Therefore, the present study used this data set.

The second author developed an Arabic BLEU System to compute the BLUE score. We have found out in Table 1 that the overall translation precision for Google Translate system is 0.449625, and the overall translation precision for the Babylon MT system is0.39845using BLEU method. On the other hand, when METEOR 1.5 [35] is used we found in Table 2, that the overall translation precision for Google Translate system is 0.3715 ((0.415 + 0.328)/2), and the overall translation precision for the Babylon MT system is 0.3205 ((0.343+0.298)/2).

The second author developed an Arabic BLEU System to compute the BLUE score. Blue method results are presented in Table 1, and it clearly shows that the overall translation precision for Google Translate MT system is approximately 0.45, and the overall translation precision for the Babylon MT system is approximately 0.4. Therefore, it is deduced that the translation accuracy from Arabic into English of Google Translate MT system is more accurate than its counterpart. On the other hand, when METEOR 1.5 [34] is used, it is found, in Table 2, that the overall translation precision for Google Translate system is 0.3715 ((0.415 + 0.328)/2), and the overall

translation precision for the Babylon MT system is 0.3205 ((0.343+0.298)/2). Once again, it is deduced that the translation accuracy from Arabic into English of Google Translate MT system is more accurate than its counterpart.

Babylon MT system proves to be more effective in translating exclamatory Arabic sentences to English.

Furthermore, it is concluded that BLEU method is closer to human evaluation than METEOR 1.5 for the translation from Arabic into English. This unexpected result is due to the fact that version METEOR 1.5, which does not fully support the Arabic language, is used in this study.

VI. Future Work

As future work, we would like to extend the scope of the study, by using a larger data set of sentences, use more automatic evaluation MT methods like ROUGE, NIST and RED, and use more languages.

Furthermore, we plan to study the effect of using an Arabic Stemmer like Khoja Stemmer [35] on the results of METEOR method.

References

[1] J. Giménez, and L. Màrquez, "Linguistic measures for automatic machine translation evaluation,"*Machine Translation* vol.24, no. 3-4, pp. 209-240, 2010.

[2] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "*BLEU: a method for automatic evaluation of machine translation*," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Stroudsburg, PA, USA, pp. 311-318, 2002.

[3] A. Lavie, K. Sagae, and S. Jayaraman, "*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 (AMTA-2004), pp. 134–143, Washington, DC, September, 2004.

[4] S. Banerjee, and A. Lavie, "*METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*" in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005.

[5] Arabic language - *Wikipedia*, the free encyclopedia. [cited 06 July 2015]. Available from:http://en.wikipedia.org/wiki/Arabic_language

[6] K. C. Ryding, "*A Reference Grammar of Modern Standard Arabic*," Cambridge: Cambridge University Press, 2005.

[7] M. Carpuat, Y. Marton, and N. Habash, "*Improving Arabic-to-English Statistical Machine Translation by Reordering Post-verbal Subjects for Alignment*," in Proceedings of the ACL 2010 Conference Short Papers, pp. 178–183, Uppsala, Sweden, 2010.

[8] R. Al Dam, and A. Guessoum, "*Building a neural network-based English-to-Arabic transfer module from an unrestricted domain*," In Proceedings of IEEE International Conference on Machine and Web Intelligence (ICMWI), pp.94-101, 2010.

[9] J. Riesa, B. Mohit, K. Knight, and D. Marcu, "*Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources*", In the Proceedings of INTERSPEECH, Pittsburgh, USA, 2006.

[10] Adly, N. and Alansary, S. 2009, "*Evaluation of Arabic Machine Translation System based on the Universal Networking Language*," in Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems "NLDB 2009", pp. 243-257, 2009.

[11] Y. Salem, A. Hensman, and B. Nolan, "Towards Arabic to English Machine Translation,"*ITB Journal*, Issue 17, pp. 20-31, 2008.

[12] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Evaluating Arabic to English Machine Translation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 5, No. 11, pp. 68 - 73, 2014.

[13] K. Kirchhoff, D. Capurro,and A. M. Turner,"A conjoint analysis framework for evaluating user preferences in machine translation,"*Machine Translation*, vol.28, no. 1,pp. 1-17, 2014.

[14] A. Lavie, and A. Agarwal, "*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, Prague, pp. 228–231, June 2007.

[15] A. Lavie, M. J. Denkowski, "The Meteor metric for automatic evaluation of machine translation," *Machine Translation*, September 2009, Volume 23, Issue 2-3, pp. 105-115, 2009.

[16] M. Denkowski, and A. Lavie, "*Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level*," Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, pp. 250-253, 2010.

[17] M. Denkowski and A. Lavie, "*Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems*," in Proceedings of the 6th Workshop on Statistical Machine translation of the Association for Computational Linguistics (ACL-WMT '11), pp. 85–91, ACL Press, 2011.

[18] B. Babych, and A. Hartley, "*Extending the BLEU MT evaluation method with frequency weightings*," In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04). Association for Computational Linguistics, Stroudsburg, PA, USA, Article 621, 2004.

[19] M. Yang, J. Zhu, J. Li., L. Wang, H. Qi, S. Li., and L. Daxin, "*Extending BLEU Evaluation Method with Linguistic Weight*," 2008. ICYCS 2008. The 9th International Conference for Young Computer Scientists, pp.1683-1688, 2008.

[20] B. Chen, and R. Kuhn, "*AMBER : A modified BLEU, enhanced ranking metric*,"in Proceedings of the 6th Workshop on Statistical Machine Translation, Edinburgh, UK, pp. 71-77, 2011.

[21] A. Guessoum, and R. Zantout, "A Methodology for Evaluating Arabic Machine Translation Systems,"*Machine Translation*, issue 18, pp. 299-335, 2005.

[22] H. Al-Haj, and A. Lavie A., "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation,"*Machine Translation*, vol. 26, no. 1-2, pp. 3-24, 2012.

[23] D. D. Palmer, "*User-centered evaluation for machine translation of spoken language*," in Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.5, pp. v/1013- v/1016, 2005.

[24] S. Condon, M. Arehart, D. Parvaz, G. Sanders, C. Doran, and J. Aberdeen, "Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics", *Machine Translation*, vol. 26, Nos. 1-2, pp. 159-176, 2012.

[25] S. Alansary, M. Nagi, and N. Adly, "*The Universal Networking Language in Action in English-Arabic Machine Translation*," In Proceedings of 9th Egyptian Society of Language Engineering Conference on Language Engineering, (ESOLEC 2009), Cairo, Egypt, 2009.

[26] S. Alansary, "*Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas*", in Proceedings of the 11th International Conference on Language Engineering, Cairo, Egypt, 2011.

[27] A. Alqudsi, N. Omar, and K. Shaker, "Arabic Machine Translation: a Survey", *Artificial Intelligence Review*, pp.1-24, 2012.

[28] T. Hailat,, M. N. AL-Kabi, I. M. Alsmadi, E. Shawakfa, "*Evaluating English To Arabic Machine Translators*," 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT 2013) - IT Applications & Systems, Amman, Jordan, 2013.

[29] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating English to Arabic Machine Translation Using BLEU,"*International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, no.1, pp. 66-73, 2013.

[30] H. Al-Deek, E. Al-Sukhni, M. Al-Kabi, M. Haidar, "*Automatic Evaluation for Google Translate and IMTranslator Translators: An Empirical English-Arabic Translation*," The 4th International

Conference on Information and Communication Systems (ICICS 2013). ACM, Irbid, Jordan, 2013.

[31] L. S. Hadla, T. M. Hailat, M. N. Al-Kabi, "Evaluating Arabic to English Machine Translation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, SAI Publisher, 5(11), pp. 68-73, 2014.

[32] A. Agarwal and A. Lavie, "*Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output*," in Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, pp. 115–118, 2008.

[33] *Meteor 1.5: Automatic Machine Translation Evaluation System, Code* by Michael Denkowski WebsiteGithub. [cited August 31, 2015]. Available from: http://www.cs.cmu.edu/~alavie/METEOR/README.html.

[34] Alon Lavie and Abhaya 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 (StatMT '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 228-231.

[35] M. Denkowski, A. Lavie, "*Meteor: Automatic Machine Translation Evaluation System*." [cited August 31, 2015]. Available from:http://www.cs.cmu.edu/~alavie/METEOR/

[36] Shereen Khoja - *Research*. [cited August 31, 2015]. Available from: http://zeus.cs.pacificu.edu/ shereen/research.htm

AUTHORS PROFILE

Laith Salman Hassan Hadla, born in Baghdad/Iraq in 1970. He obtained his PhD in Machine Translation from Al-Mustansiriya University in (2006), his masters' degree was in stylistic translation from Al-Mustansiriya University in (1995), and his bachelor degree is in Translation from Al-Mustansiriya University in (1992). Laith S. Hadla is an assistant professor at the Faculty of Arts, at Zarqa University. Before joining Zarqa University, he worked since 1993 in many Iraqi and Arab universities. The main research areas of interest for Laith S. Hadla are machine translation, translation in general, and linguistics. His teaching interests fall into translation and linguistics

Taghreed M. Hailat, born in Irbid/jordan in 1986. She obtained her MSc. degree in Computer Science from Yarmouk University (2012), and her bachelor degree in Computer Science from Yarmouk University (2008). Currently, she is working at Irbid chamber of Commerce as a Computer Administrator and previously as a trainer of many computer courses at Irbid Training Center

Mohammed Al-Kabi, born in Baghdad/Iraq in 1959. He obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his master's degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq(1981). Mohammed Naji Al-Kabi is an assistant Professor in the Faculty of IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a lecturer in PSUT and Jordan University of Science and Technology (JUST). Al-Kabi's research interests include Information Retrieval, Sentiment analysis and Opinion Mining, Big Data, Web search engines, Machine Translation, Data Mining, & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Big Data, Web programming, data mining, DBMS (ORACLE & MS Access).