Evaluating English to Arabic Machine Translation Using BLEU

Mohammed N. Al-Kabi
Faculty of Sciences & IT
Zarqa University
Zarqa – Jordan

Taghreed M. Hailat, Emad M. Al-Shawakfa, and Izzat M. Alsmadi
Faculty of IT and CS
Yarmouk University,
Irbid 211-63, Jordan.

***Abstract*— This study aims to compare the effectiveness of two popular machine translation systems (Google Translate and Babylon machine translation system) used to translate English sentences into Arabic relative to the effectiveness of English to Arabic human translation. There are many automatic methods used to evaluate different machine translators, one of these methods; Bilingual Evaluation Understudy (BLEU) method, which was adopted and implemented to achieve the main goal of this study. BLEU method is based on the assumptions of automated measures that depend on matching machine translators' output to human reference translations; the higher the score, the closer the translation to the human translation will be. Well known English sayings in addition to manually collected sentences from different Internet web sites were used for evaluation purposes. The results of this study have showed that Google machine translation system is better than Babylon machine translation system in terms of precision of translation from English to Arabic.**

Keywords*— component; Machine Translation; Arabic; Google Translator; Babylon Translator; BLEU

I. INTRODUCTION

Machine translation is a task that involves the process of translating a source sentence from one language giving the meaning into another target language(s). Online machine translators rely on different approaches to translate from one natural language into another, these approaches are: Rule-based, Direct, Interlingua, Transfer, Statistical, Example-based, Knowledge-based, and Hybrid Machine Translation (MT).

The accuracy of any machine translator is usually evaluated by comparing the results to human judgments. One of the methods used to evaluate machine translation systems is called BiLingual Evaluation Understudy (BLEU) which was introduced in the study of Papineni, Roukos, Ward, and Zhu [1] and claimed to be language independent and highly correlated with human evaluation.

BLEU is based on a core idea to determine the quality of any machine translation system which is summarized by the closeness of the candidate output of the machine translation system to reference (professional human) translation of the same text.

The closeness of the candidate translation to the reference translation is determined by a modified n-gram precision which was proposed by Papineni, Roukos, Ward, and Zhu [1]. The modified n-gram precision is the main metric adopted by BLEU to distinguish between good and bad candidate translations, where this metric is based on counting the number of common words in the candidate translation and the reference translation, and then divides the number of common words by the total number of words in the candidate translation. The modified n-gram precision penalizes candidate sentences found shorter than their reference counter parts, also it penalize candidate sentences which have over generated correct word forms.

The US National Institute of Standards and Technology (NIST) have presented a new method called NIST; which represents an enhancement to BLEU. The NIST method is used to evaluate the effectiveness of a number of machine translation systems to translate from various natural languages into English. This method, and according to Doddington [2], tries to compute how informative a particular n-gram is, where a low frequency of a particular n-gram means yielding a higher weight, while a high frequency of a particular n-gram means yielding a lower weight.

Due to its rich and complex morphological features, Arabic has always been a challenge for machine translation. In addition, Arabic has different word forms and word orders which make it possible to express any sentence in different forms.

Furthermore, the existence of many dialects and the fact that the word order is not usually the same for source and target languages, this leads usually to the possibility of having more than one meaning for the same sentence according to Alqudsi, Omar, and Shaker [21]. English-to-Arabic machine translation has been a challenging research issue for many of the researchers in the field of Arabic Natural Language Processing.

Many attempts were made to perform or enhance machine translation of Arabic into other languages. Some of these attempts are the work of Al Dam, and Guessoum [3], Carpuat, Marton, and Habash [4], Adly and Al-Ansary [5], Salem, Hensman, and Nolan [6], and Riesa, Mohit, Knight, and Marcu [7].

To evaluate any translation system, one should use a proper corpus. As for Arabic, the authors could not find any standard corpus that could be used for evaluation purposes.

For this, we had to collect our data from different Internet websites representing two types of datasets; a set of well-known English sayings and a set of sentences that were translated manually by two human translators for judgment purposes. In this study, we have evaluated the effectiveness of two automatic machine translators that could be used for English-to-Arabic translation and vice versa. The used machine translators are Google machine translator and the Babylon machine translator.

This paper is organized as follows: section 2 presents the related work, section 3 presents the methodology followed in this research, section 4 presents the evaluation of machine translators under consideration, through the usage of a system designed and implemented by one of the authors. Section 5 presents the conclusion from this research, and last but not least section 6 discusses extensions of the this study and the future plans to improve it.

II. Literature Review

A number of studies were conducted to evaluate the translation quality using an automated tool. One of these researches was conducted by Nießen, Och, Leusch, and Ney [8]. In their study, they have presented the typical necessary requirements to build an effective tool for the evaluation of the accuracy of different machine translators. Word Error Rate (WER) and Subjective Sentence Error Rate (SSER) are discussed as two essential criteria to the quality of the outputs of machine translators. The authors have described their technique as fast, semiautomatic, convenient and consistent.

Precision, Recall, and F-measure are famous measures which are usually used to evaluate information retrieval systems (IRS) and search engines, however, in their study Melamed, Green, and Turian [9], have showed that these three measures can also be used to evaluate machine translators, and further showed that these three measures are highly correlated to these measures. In addition, the authors claimed that these measures are more reliable than Bilingual Evaluation Understudy (BLEU).

Usually, Machine translation evaluation methods are based on reference translations, but that is not always the case. So for example, Palmer [10] has introduced a user-centered method in his study to evaluate machine translation systems that is based on comparing the outputs of machine translation systems and then ranked, according to their quality, by expert users who have the necessary needed scientific and linguistic backgrounds to accomplish the ranking process. His study covers four Arabic-to-English and three Mandarin (simplified Chinese)-to-English machine translation systems.

Another method for evaluating machine translation systems was presented by Akiba et al. [11]. Their study was dedicated to evaluate machine translation (MT) systems that are subsystems of speech-to-speech MT (SSMT) systems. The researchers referred to the two drawbacks of using BLEU to evaluate SSMT, where the first drawback was related to the position based error assessment, while the second drawback was related to the tolerance to accept colloquial sentences. The new method presented in their paper was called “gRader based on Edit Distances (RED)”, which automatically computes the score related to the translated output of the machine translation system using a decision tree (DT). They have conducted a series of experiments which revealed; according to the authors, that their novel method RED is more accurate than BLEU method.

The BLEU method is characterized by the fact that it is language independent and not designed for a certain natural language. BLEU has a number of cons, therefore a number of researchers have attempted to enhance this important method. One of such attempts was conducted by Yang et al. [12]. They have used linguistic features of the evaluated sentences outputted from the machine translation systems in their enhancements. Those researchers have used multiple linear regressions to assign proper weights to different n-grams and words within BLEU framework. These enhancements helped in improving the effectiveness of the BLEU method.

Both BLEU and NIST are widely used metrics to evaluate machine translation systems' outputs. Since they are language independent, these two methods ignore the linguistic features of the targeted natural language. A study by Qin, Wen, and Wang [13] have noticed this fact and thus used synonymous words and phrases to those found in the reference translations. In their study, a N-gram co-occurrence algorithm was used to produce pseudo translations for BLEU and NIST, the pseudo translations are based on substituting words and phrases in the reference translations for synonyms. Tests on this method have revealed clearly that the enhancement to both BLEU and NIST is more correlated to human evaluations.

In their study, Veillard, Melissa, Theodora, Racoceanu, and Bressan [14] have adopted machine learning (ML) to evaluate machine translation (MT) systems, and have proposed a new ML-based metrics, which uses support vector machine methods and include multi-class support vector machines (SVM) and support vector regression (SVR) with different kernel functions. Tests on these new ML-based metrics proved that they outperform the popular standard metrics like BLEU, METEOR, and ROUGE.

Most of the previous studies presented in this study are related to an automatic evaluation of machine translation on sentence level, where the connectivity of sentences in a document is neglected. Wong, Pun, Kit, and Webster [15] study however, is characterized by presenting a new metric to automatically evaluate the quality of the translation at a document level. They have emphasized on the structure of the outputted document by the machine translation system, more specifically, on the lexical cohesion feature. Conducted tests by the researchers showed that the adopted feature is influential and helps to improve the correlation between human judgments of machine translation outputs at the document level by 3% to 5%.

Brkic, Mikulic, and Matetic [16] have conducted a study to evaluate the machine translation (MT) from Croatian to English using two MT systems (Google Translate) and a system called LegTran that was developed and introduced by her. A reference translation conducted by a professional translator is also used. WER, PER, TER, F-measure, BLEU, and NIST as automatic evaluation methods were used. The conducted tests showed that there is no contradiction between

the results of the above six methods used to identify the best MT system, except that human BLEU scores were higher than the automated BLEU score.

Condon et al. [17] study was related to the automatic evaluation of Iraqi Arabic–English speech translation dialogues. Those researchers have found that translation into Iraqi Arabic will correlate higher with human judgments when normalization (light stemming, lexical normalization, and Orthographic normalization) is used.

In their study, Adly and Al-Ansary [5] have conducted an evaluation of Arabic machine translation based on the Universal Networking Language (UNL) and the Interlingua approach for translation. The Interlingua approach relies on transforming text in the specified language into a representation form that is language independent that can be later on transferred into the target language. Three measures were used for the evaluation process; BLEU, F1 and Fmean. The evaluation was performed using the Encyclopedia of Life Support Systems (EOLSS). The effect of UNL onto translation from/into Arabic language was also studied by Alansary, Nagi, and Adly [18], and Al-Ansary [19]

The different characteristics of the Arabic language and their effect on Machine Translation were the topic of Salem, Hensman, and Nolan [6] study. In their study, the authors have proposed a model incorporating the Role and Reference Grammar technique to overcome the free word order of Arabic obstacle in the Translation process.

Carpuat, Marton, and Habash [4] study has addressed the challenges raised by the Arabic verb and subject detection and reordering in Statistical Machine Translation. To minimize ambiguities, the authors have proposed a reordering of Verb Subject (VS) construction into Subject Verb (SV) construction for alignment only which has led to an improvement in BLEU and TER scores.

A methodology for evaluating Arabic machine translation was presented in the study of Guessoum and Zantout [20]. In their study, they have evaluated lexical coverage, grammatical coverage, semantic correctness and pronoun resolution correctness. Their approach was used to evaluate four English-Arabic commercial Machine Translation systems; namely ATA, Arabtrans, Ajeeb, and Al-Nakel.

In a recent survey by Alqudsi, Omar, and Shaker [21], the issue of machine translation of Arabic into other languages was discussed. In the survey, the challenges and features of Arabic for machine translation was discussed. In addition, different approaches to machine translation and their possible application for Arabic were also mentioned in the survey. The survey concluded by indicating the difficulty of finding a suitable machine translator that could meet human requirements.

In a study by Galley, Green, Cer, Chang, and Manning [22], an Arabic-to-English statistical machine translator called the Stanford University's Arabic-to-English SMT which was built as in improvement to a previous Chinese-to-Arabic MT system was described. In their system, a comparison between three types of lexicalized reordering models was performed.

A phrase-based reordering model was used as the core engine of the system and the BLEU score was reported to have increased using their approach.

In a study by Khemakhem, Jamoussi, and Ben Hamadou [23], an Arabic-English Statistical Machine translator; called MIRCL, was discussed. The MIRCL system was built using a phrase-based approach. In addition, a solution for disambiguation of the output of the Arabic morphological analyzer was presented in their study that was used to help in selecting the proper word segments for translation purposes.

The impact of Arabic morphological segmentation on the performance of a broad-coverage English-to-Arabic Statistical machine translation was discussed in the work of Al-Haj and Lavie [24]. In their work, a phrase based statistical machine translation was addressed. Their results have showed a difference in BLEU scores between the best and worst morphological segmentation schemes where the proper choice of segmentation has a significant effect on the performance of the SMT.

III. The Methodology

This section presents the main steps, followed to accomplish this study, and summarized in Figure 1. Bilingual Evaluation Understudy (BLEU) method is adopted in this study to evaluate Babylon machine translation system and Google Translate machine translation system. The effectiveness of translation from English to Arabic using Babylon machine translation system and Google Translate system is tested using BLEU method.

In the first step we have to input 5 statements as shown below:

The source sentence in English is inputted to the machine translation system.

The translation of the source sentence using Google Translate system.

The translation of the source sentence using Babylon Translate system.

Two reference translations of the source sentence.

The second step involves the text preprocessing by dividing the text into different n-gram sizes, as follows: unigrams, bigrams, trigrams, and tetra-grams. The precision for Babylon machine translation system and Google machine translation system were computed for each of the four gram sizes. In the final step, for each of the four n-gram sizes, we compute a unified precision score for that size. These values are then compared to decide which of them get the best translation.

A. Dividing the text into different n-gram sizes

An n-gram can be defined as a sub-sequence of n items, from a given sequence of words (text or sentence). These items can be characters, words or sentences according to the application.

An n-gram can be of any number of words and each of which has a name, when the sizes of the n-grams are equal to one, two, three, or four words, they are called unigram, bigram, trigram, and tetra-gram respectively. This study deals

with these types. The n-gram extraction technique to extract any size of word(s) is described in Figure 2

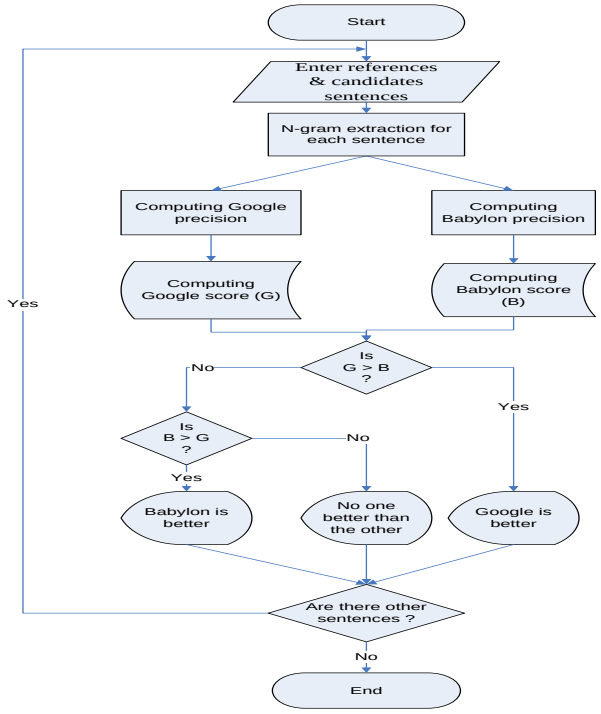

Figure 1. Evaluation Methodology Flowchart.

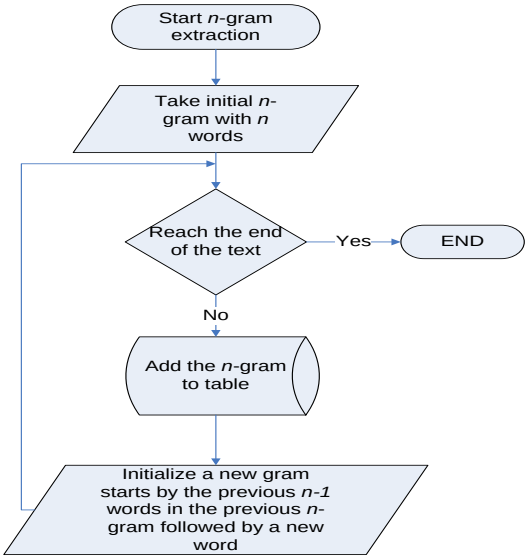

Figure 2. N-grams Extraction Flowchart

To explain this method for extracting n-grams, we will provide an example for bigram size in the study; so we translate the statement "World football cup is held every four years once" into Arabic as "يعقد كأس العالم لكرة القدم كل اربع سنوات مرة", and divide it into bigrams as shown in Figure 3 below

Figure 3. Bigram Example.

B. *Babylon and Google Machine Translators Precision*

N-grams are used in many areas like information retrieval, text mining, natural language processing (NLP) … etc. In this study, n-gram extraction is used as a preprocessing technique. In order to compute the precision score for each of the four n-gram sizes, we have to count first the number of common words in every candidate and reference sentence, and then we have to divide this sum over the total number of n-grams in the candidate sentence.

To explain that, we take a source sentence as an example and translate it using Babylon machine translation system and Google Translate machine translation system, and two human translations called Reference 1 and Reference 2 as follows.

EXAMPLE 1:

Source Sentence: Banks usually lend money to persons who need it, for a specified interest.

Babylon machine translation system:

أجل من المال يإل يـ حـ تاجون الـ ذيـ ن الأ شخاص عادة تـ قدم مـ صارف محددة مـ صالـ ح

Google Translate:

و إلـ يها، يـ حـ تاجون الـ ذيـ ن لـ لأ شخاص عادة الـ مال تـ قرض الـ بـ نوك محددة لـ مـ صـ لحة ذلـ ك

Reference 1:

الأموال يـ حـ تاجون الـ ذيـ ن الأ شخاص بـ إقـ راض الـ بـ نوك تـ قوم مـ عـ يـ نه فـ ائـ دة مـ قابـ ل

Reference 2:

لـ قاء لـ ها يـ حـ تاجون الـ تـ ي الـ مـ بالـ غ الـ ناس تـ قرض عادة الـ مـ صارف لـ لمـ صرف فـ ائـ دة

At this stage we have to compare the outputs of Google Translate system with the two references. The first comparison is based on unigram; we found that the unigrams "الـ بـ نوك Banks", "الـ ذيـ ن who" and "يـ حـ تاجون need" are common with reference 1, also "تـ قرض lend" and "يـ حـ تاجون need" with reference 2. So, the number of common unigrams is equal to 4.

The total unigrams in output of Google Translate system for the source sentence is equal to 12. So the unigram precision is equal to (4/12) ≈ 0.33, as shown in Table 1.

Then, when we do the second comparison according to bigram, we found that "يد حد تاجون الذين who need" bigram is the only bigram common with reference 1, with no bigrams common with reference 2. So the bigram precision is equal to (1/11) ≈ 0.09. The trigram precision and tetra-gram precision values were computed in the same way, and the results are shown in Table 1.

TABLE I. PRECISION VALUES FOR EXAMPLE 1.

MT / N-grams	Babylon machine translation system	Google Translate System
Uni-gram Precision(P1)	$\frac{3}{12}$	$\frac{4}{11}$
Bi-gram Precision (P2)	$\frac{1}{11}$	$\frac{1}{10}$
Tri-gram Precision(P3)	$\frac{0}{10}$	$\frac{0}{9}$
Tetra-gram Precision(P4)	$\frac{0}{9}$	$\frac{0}{8}$

C. Babylon and Google Machine Translators BLEU-score

To combine the previous precision values in a single overall score (called BLEU-score), we start by computing the Brevity Penalty (BP) by choosing the effective reference (i.e. the reference that has more common n-grams) length which is denoted by r. Then we compute the total length of the candidate translation denoted by c. Now we need to select Brevity Penalty to be a reduced exponential in (r / c) as shown in equation 1 [1]:

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

In our example for Babylon machine translation system c = 12, r = 10, and when 12 > 10 then the BP = 1, and for Google Translate c = 11, r = 10, and when 11 > 10 then also BP = 1.

Now, we use the previous resulted BP from equation 1 to compute the final BLEU score as shown in formula (2) [1].

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

where N = 4 and uniform weights wn = (1/N), in this study [1].

Tests on example1 showed that the BLEU score for the Babylon machine translation system is 0.075, and the BLEU score for the Google Translate is 0.115. This result indicates that Google Translate is more accurate than Babylon machine translation system, since higher BLEU score for any machine translator means that its better than its counterparts with lower BLEU scores.

Papineni, Roukos, Ward, and Zhu [1] study noted that the BLEU metric ranges from 0 to 1, where the translation that has a score of 1 is identical to a reference translation [1].

IV. THE EVALUATION

In order to speed up the calculation used in evaluating the Babylon machine translation system and the Google machine translation system we have developed a system using visual studio .Net 2008 to accomplish this goal, the main screen of the system is shown in Figure 4 as shown below.

As indicated by Alqudsi, Omar, and Shaker [21], most of the approaches that have been proposed for Arabic-English machine translation was tested on limited domains; mostly news and government data. For this, to evaluate the attained results of this evaluation system, we have constructed a corpus of 100 sentences that were categorized into 7 types; past, present, future, imperative, passive, conditional "if", and questions. In addition to that, 300 popular English sayings were also taken and translated into Arabic using both the Babylon and Google translators.

The majority of the conducted experiments on these sayings have resulted into a literal and meaningless translation of the saying. For instance, the English say "A good workman is known by his chips"; which has the Arabic meaning as "عند يهان أو المرؤ يد كرم الامم تحان", was literally mistranslated by both translators into "الى شرائح معروف جيد عامل", as a Babylon translation, and into "له رقائق من الجيد العامل المعروف ومن", as the Google translation; which is very literal and very far from the actual meaning of the saying.

In our evaluation and testing of the translators, we have found out that in some sentences the translation precision is equal for both machine translators (Google and Babylon). However, after the application of the Arabic BLEU system on the 300 English sayings, the conducted experiments have indicated better translation accuracy by the Google translator than the Babylon translator; (0.44 for Google and 0.12 for Babylon).

It has also been noticed that Babylon translator have not succeeded in correctly translating any of the sayings at 100% accuracy, and that Google translator have succeeded; at some extent, in fully translating some of these sayings. For general translations, it has been noticed that "Google Translate system was better than Babylon machine translation system in most of the translations".

As a whole, the average precision values of Google and Babylon machine translation system for each type of sentences in the corpus are shown in Table 2 and Figure 5. It is obvious that Google Translate system was better than Babylon machine translation system.

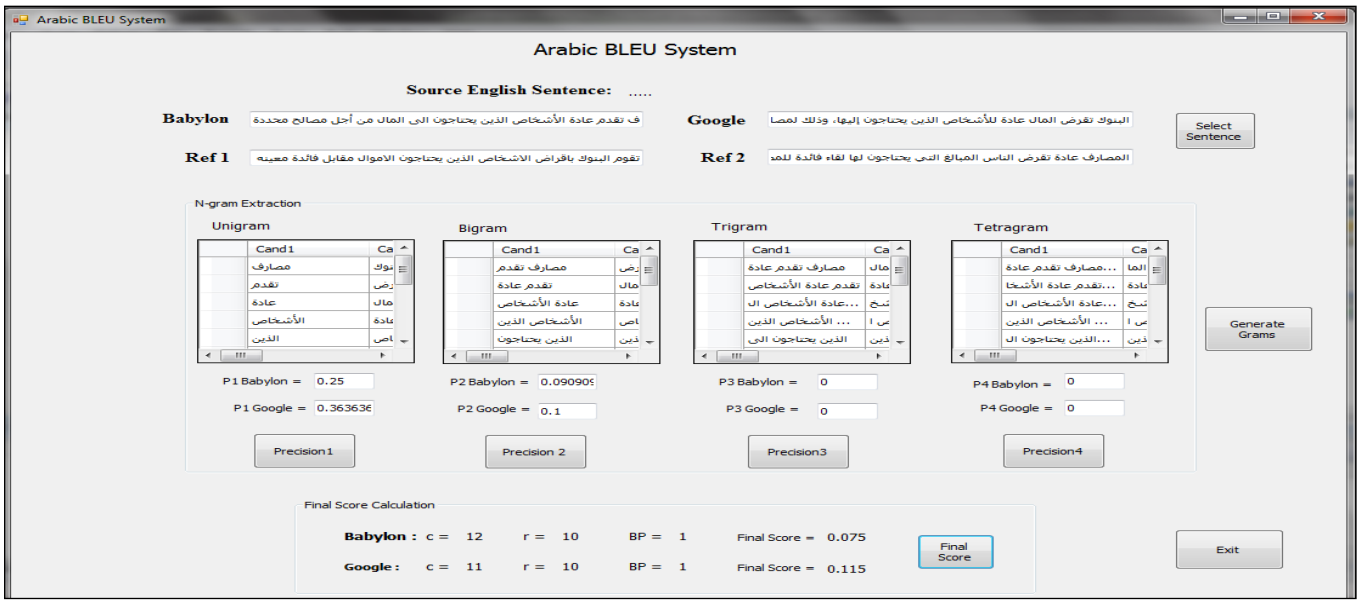

Figure 4. The Main Screen of the Arabic BLEU System.

TABLE II. AVERAGE PRECISION FOR EACH TYPE OF SENTENCES TYPE

Type / Translator	Past	Present	Future	Imperative	Passive	Conditional "if"	Questions
Babylon machine translation system	0.193	0.206	0.172	0.196	0.239	0.146	0.205
Google machine translation system	0.386	0.414	0.267	0.404	0.273	0.163	0.294

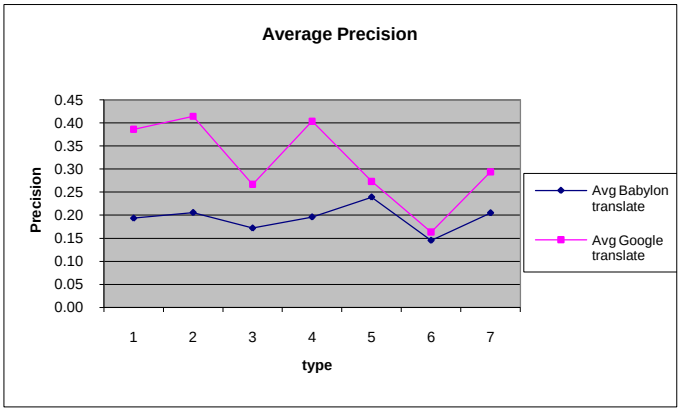

Figure 5: Summary of Average Precision

We have noticed from the conducted experiments that the translation quality of 21% of translated English sentences into Arabic using Babylon translate system were more accurate than Arabic sentences outputted by Google Translate system. While the translation quality of Google Translate system was better than the translation quality of Babylon translate system to translate 69% English sentences into Arabic. The two machine translators yield equal accuracy to translate 10% of the English sentences into Arabic.

V. CONCLUSION

English-to-Arabic machine translation has been a challenging research issue for many of the researchers in the field of Arabic Natural Language Processing. In this study, we have evaluated the effectiveness of two automatic machine translators that could be used for English-to-Arabic translation and vice versa. The used machine translators are Google machine translator and the Babylon machine translator.

The accuracy of any machine translator is usually evaluated by comparing the results to human judgments. There is no standard Arabic corpus that can be used for such evaluations, for this we had to collect our data from different Internet websites representing two types of data; a set of well-known English sayings and a set of sentences that were translated manually by two human translators for judgment purposes.

Although the collected data was of small size, the well-known English sayings usually presented a challenge for the Machine translators into Arabic.

After applying our developed Arabic BLEU System on the collected data, we have found out that the overall translation precision for Google was 0.314 and the overall translation precision for the Babylon machine translation system was 0.194. As for the English popular sayings, it has been found out that the Google translate system has better accuracy than that of Babylon translation system (0.44 for Google and 0.12 for Babylon). Based on these findings, we can conclude that the Google Translate system is better than Babylon machine translation system for the translation from English into Arabic.

Furthermore, we have found out that Babylon machine translation system was incapable of translating some of the English words into Arabic properly. For example, the Babylon machine translator could not fully translate the following English sentence: "Great talkers are little doers ", since the outputted Arabic translation was: "يراقب top talkers كبيرة الدين قليلة".

VI. Future Work

Measures of translation quality based on exact matching of word forms are of challenge because of the orthographic variation; which is especially severe in the Arabic language. To solve such problem, and as a future research, we are planning to find a technique to solve it. Other automatic evaluation methods for machine translators like NIST, METEOR, ROUGE and RED will be included in our future studies.

We have tested our experiments on a small size of data, as part of the future work we are planning on collecting more data and perform tests using the new data as well as any available standard data that could be found.

References

[1] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. "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.

[2] Doddington G. 2002. "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". In Proceedings of the second international conference on Human Language Technology Research (HLT '02). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 138-145.

[3] Al Dam, R.; Guessoum, A. 2010. "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.

[4] Carpuat M., Marton Y., and Habash N., 2010. "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.

[5] 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.

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

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

[8] Nießen S., Och F.J., Leusch G., Ney H.. 2000. "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research". In Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 39-45.

[9] Melamed D., Green R., and Turian J.P., 2003. "Precision and recall of machine translation". In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers - Volume 2 (NAACL-Short '03), pp. 61-63.

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

[11] Akiba, Y.; Imamura, K.; Sumita, E.; Nakaiwa, H.; Yamamoto, S.; Okuno, H.G.; 2006, "Using multiple edit distances to automatically grade outputs from Machine translation systems," Audio, Speech, and Language Processing, IEEE Transactions on , vol.14, no.2, pp. 393- 402.

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

[13] Qin Y., Wen Q., Wang J., 2009. "Automatic evaluation of translation quality using expanded N-gram co-occurrence,", NLP-KE 2009. International Conference on Natural Language Processing and Knowledge Engineering, pp.1-5.

[14] Veillard, A.; Melissa, E.; Theodora, C.; Racoceanu, D.; Bressan, S. 2010. "Support Vector Methods for Sentence Level Machine Translation Evaluation,", 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), vol.2, pp.347-348.

[15] Wong, B.T.M., Pun, C.F.K., Kit, C., Webster, J.J. 2011. "Lexical cohesion for evaluation of machine translation at document level," 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), 2011, pp.238-242.

[16] Brkic, M. Mikulic, B.B. ; Matetic, M. ; Basic Mikulic, Bozena; Matetic, Maja; 2012. "Can we beat Google Translate?,", Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI) , pp.381-386.

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

[18] Alansary S., Nagi M. and Adly N. 2009. "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.

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

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

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

[22] Galley M., Green S., Cer D., Chang P.C., and Manning C.D. 2010, " Stanford University's Arabic-to-English Statistical Machine Translation System for the2009 NIST MT Open Evaluation", in NIST Open Machine Translation Evaluation Meeting.

[23] Khemakhem I., Jamoussi S., Ben Hamadou A., 2010, "The MIRACL Arabic-English Statistical Machine Translation System for IWSLT 2010", in Proceedings of the 7th International Workshop on Spoken Language Translation, Paris, France.

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

AUTHORS PROFILE

Mohammed Al-Kabi 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 masters 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 Sciences and 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 Jordan University of Science and Technology. AL-Kabi's research interests include Information Retrieval, Web search engines, Data Mining, Software Engineering & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Web programming, data mining, DBMS (ORACLE & MS Access).

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.

Emad M. Al-Shawakfa is an Assistant Professor at the Computer Information Systems Department at Yarmouk University since September 2000. He was born in Jordan in 1964 and holds a PhD degree in Computer Science from Illinois Institute of Technology (IIT) – Chicago, USA in the year 2000, a MSc in Computer Engineering from Middle East Technical University in Ankara-Turkey in the year 1989, and a BSc in Computer Science from Yarmouk University in Irbid-Jordan in the year 1986. His research interests are in Computer Networks, Data Mining, Information Retrieval, and Arabic Natural Language Processing. He has several publications in these fields and currently working on others

Izzat M. Alsmadi. is an associate professor in the CIS department at Yarmouk University, Irbid, Jordan. Born in Jordan 1972, Izzat Alsmadi had his master and PhD in software engineering from North Dakota State University (NDSU), Fargo , USA in the years 2006 and 2008 respectively. His main areas of research include: software engineering, testing, metrics, and information retrieval**.**