Evaluating Urdu to Arabic Machine Translation Tools

Maheen Akhter Ayesha[1], Sahar Noor[2], Muhammad Ramzan[3], Hikmat Ullah Khan[4], Muhammad Shoaib[5]
[1,2,5]Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal, Pakistan
[3]Department of CS & IT, University of Sargodha, Sargodha, Pakistan
[4]Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan

***Abstract*—Machine translation is an active research domain in fields of artificial intelligence. The relevant literature presents a number of machine translation approaches for the translation of different languages. Urdu is the national language of Pakistan while Arabic is a major language in almost 20 different countries of the world comprising almost 450 million people. To the best of our knowledge, there is no published research work presenting any method on machine translation from Urdu to Arabic, however, some online machine translation systems like Google[1], Bing[2] and Babylon[3] provide Urdu to Arabic machine translation facility. In this paper, we compare the performance of online machine translation systems. The input in Urdu language is translated by the systems and the output in Arabic is compared with the ground truth data of Arabic reference sentences. The comparative analysis evaluates the systems by three performance evaluation measures: BLEU (BiLingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering) and NIST (National Institute of Standard and Technology) with the help of a standard corpus. The results show that Google translator is far better than Bing and Babylon translators. It outperforms, on the average, Babylon by 28.55% and Bing by 15.74%.**

Keywords*—Natural language processing; machine translation; Urdu-Arabic Corpus; Google; Bing; Babylon; translator; BiLingual Evaluation Understudy (BLEU); National Institute of Standard and Technology (NIST); Metric for Evaluation of Translation with Explicit ORdering (METEOR)

I. INTRODUCTION

Urdu is the national language of Pakistan while Arabic is a major language in almost 20 different countries of the world comprising almost 450 million people. Among 7,105 languages spoken in different areas of the world, Urdu is ranked at 19th number.[4] In Pakistan, Urdu language is the medium of instruction in most of the public and private institutions. The main information sources such as newspapers and electronic media use Urdu language [1]. Arabic is the main language in 20 different countries like Egypt, Iraq, Saudi Arabia, Somalia, Sudan, Syria and the United Arab Emirates [2]. Arabic is also considered as a religious language of Muslims, as the Holy Quran and Hadith books are written in Arabic language.

Machine Translation (MT) is a process of translating a given input or source sentence from one language to the other target language. Now-a-days MT plays a significant role in different areas like education, business, medical and trade, etc. Different MT techniques such as Rule-based [3], [4], Direct [6], Transfer [5], Statistical [6], Interlingua [7], Example based [8], Knowledge-base [9] and Hybrid Machine Translation [10], [11] (MT) are used to translate from one language to the other.

All the approaches have their own pros and cons. No MT approach is the perfect in all scenarios and for all languages [12]. In this paper, we use the terms "translator", "MT system" and "MT tool" interchangeably.

A. Motivation

Pakistani and Arab communities have many things in common like cultural heritage, religion, traditions, etc. These communities need to understand each other for many reasons. A large community of Pakistani people works in Arab countries. Every year, a large number of Pakistani people travels to Arab countries to visit sacred places (Makkah, Madina), to get jobs and to promote their trade and businesses. The Arab people also visit Pakistan to get higher education and to promote their businesses. These communities need to understand each other, but there is a language barrier. Machine translation systems can help them remove this barrier. The performance of online MT systems differs a lot. A user of these MT systems may not know the best one. We, in this paper, evaluate the performance of three online MT systems to help the Arab and Pakistani communities to select the best MT system.

B. Problem

Many MT approaches have been proposed in literature for the translation of different languages. In the relevant literature, we could not find any published machine translation approach from Urdu to Arabic however some commercial machine translation systems like Google, Bing and Babylon provide Urdu to Arabic translation. The users of these translators, while translating from Urdu to Arabic, do not know the quality (accuracy level) of their translations. The users may be interested to use the best translator but they might not know the best one.

C. Contribution

In this work, we compare three online MT systems (Google, Bing and Babylon). We evaluate these MT systems by three different evaluation measures BLEU [13], METEOR [14] and NIST.[5] The results show that Google translator is better than Bing and Babylon translators. To the best of our

[1] https://translate.google.com/
[2] http://www.bing.com/translator
[3] http://translation.babylon-software.com/
[4] http://www. ethnologue. com
[5] https://www.nist.gov/

knowledge, our work is unique and the first instance of comparing the Urdu to Arabic MT systems.

Rest of the paper is organized as: Section 2 reviews related work; Section 3 formulates the problem; Section 4 describes the research methodology used for evaluation; Section 5 presents and discusses the results achieved and Section 6 provides summary and potential future work.

II. Related Work

In literature, human, automatic and embedded evaluations are three main types that are used to evaluate MT systems [21]. Many automatic techniques like BLEU, NIST and METEOR are used to evaluate the output of the MT systems. BLEU and NIST techniques overlook the linguistic characteristics of the targeted natural language because both are language independent. Ying et al. in [22] use phrases and identical words that are found in reference translation. An N-gram co-occurrence algorithm is used in their study for producing virtual translations in both techniques. METEOR uses a score based computation in finding similar words between the output of any machine translator and the reference translation given to it. Lavie et al. [23] research shows that the evaluation based on recall used in METEOR having more consistency as compared to that of precision.

As mentioned earlier, there is no research work which targets the content to be translated from Urdu to Arabic therefore we here review some research works which are related to Urdu or Arabic but the translation is aimed for other languages. Different comparative studies of MT systems from Urdu to other languages and vice versa are available in the literature [15]. Same is the case of comparative studies of MT systems from Arabic to other languages and vice versa [11], [16]-[18].

Kit and Wong [16] compare five translators (Google, PROMT[6], SDL[7], SYSTRAN[8] and WorldLingo[9]) using 13 languages (Arabic, Chinese, Dutch, French, German, Greek, Italian, Japanese, Korean, Portuguese, Russian, Spanish and Swedish) with BLEU and NIST scores. They use two reference texts i.e., Universal Declaration of Human Rights and European Union's Treaties. According to their report, SYSTRAN is the best for many languages, especially from Greek and Russian to English translation, whereas Google translator is the best in Arabic and Chinese to English translation. PROMT works better from Portuguese to German, and WorldLingo from Swedish to English than others.

Aiken and Wong [19] compare four translators (SYSTRAN, SDL, WorldLingo and InterTran[10]) using 20 Spanish phrases from an introductory textbook into English. They use human evaluators as reference translation and manually compare the translator results. According to their report, SYSTRAN and WorldLingo are better than SDL and InterTran. Vanjani et al. [20] compare SYSTRAN translator with an expert and intermediate human translator using 10 English sentences. According to their report, the fluent human translator accuracy is 100% and other's 80%. Whereas SYSTRAN got only 70% accuracy while it is faster than human by 195 times.

For Arabic to English MT, Hadla et al. [11] present the comparison of Google and Babylon Translators. The Arabic sentences are categorized in four basic sentences: imperative, declarative, exclamatory and interrogative. They report that Google translator outperforms Babylon translator. Their work is close to ours'. We perform comparative study of MT systems from Urdu to Arabic and they compare the MT systems from Arabic to English.

III. Problem Statement

There are few commercial translators that provide this translation. The users of these translators need to know the accuracy level of these translators. If it is known the users will prefer the best translator.

We formally define our problem as: *"Given the set of Urdu sentences as input to three machine translation systems, compare the output of these translators (Arabic sentences) by using multiple evaluation methods."*

Research Question: Which machine translation system is the best out of the three translators?

IV. Methodology

We compare three online machine translation systems (Google, Bing and Babylon). We use Urdu sentences as input while Arabic is output of the MT systems. The output is compared with the corresponding reference sentences (Arabic). The reference sentences are the true values or ground truth as they are manually translated by the language experts. Fig. 1 depicts the framework of the proposed methodology.

In the following subsections, we describe the corpus and the evaluation methods used in this work.

A. Corpus

We use the corpus[11] exploited by Kabi, et al. [17]. The original corpus contains Arabic and corresponding English sentences. We use all the Arabic sentences available in that corpus and corresponding Urdu sentences. We amended the original corpus by manually translating the Arabic sentences into Urdu sentences. Our corpus[12] comprises of 159 Urdu and Arabic sentences of three different types. The summary of the corpus is shown in TABLE I. We use Urdu sentences as input to the translators and, the human translated sentences (Arabic) in as reference sentences. The reference sentences are used to compare the output sentences of the MT translation systems.

The reference sentences are considered to be correct as they are generated by human experts.

[6] http://www.online-translator.com/
[7] https://www.freetranslation.com/
[8] http://www.systranet.com/translate
[9] http://www.worldlingo.com/en/products_services/worldlingo_translator.html
[10] http://transdict.com/translators/intertran.html

[11] https://docs.google.com/spreadsheets/d/1bqknBcdQ7cXOKtYLhVP7YHbvrlyJlsQggL60pnLpZfA/edit#gid=1057233962
[12] https://drive.google.com/open?id=0B-gV0w2HFyC1NlRiUkIzV3F2UUU

TABLE II shows one sample sentence of each type in Urdu and Arabic.

To evaluate the score of the corpus we use different techniques which are discussed in Performance Measures section.

V. PERFORMANCE MEASURES

We exploit three evaluation measures (BLEU, METEOR and NIST) to compare the performance (accuracy) of the three translators from Urdu to Arabic. As a rule, a machine translation that is closer to the reference translation is considered to be more accurate. This is the gist behind the machine translation evaluation methods.

TABLE I. THE CORPUS STATISTIC

Sentence Type	No. of Sentences
Declarative	70
Exclamatory	49
Imperative	40
Total	159

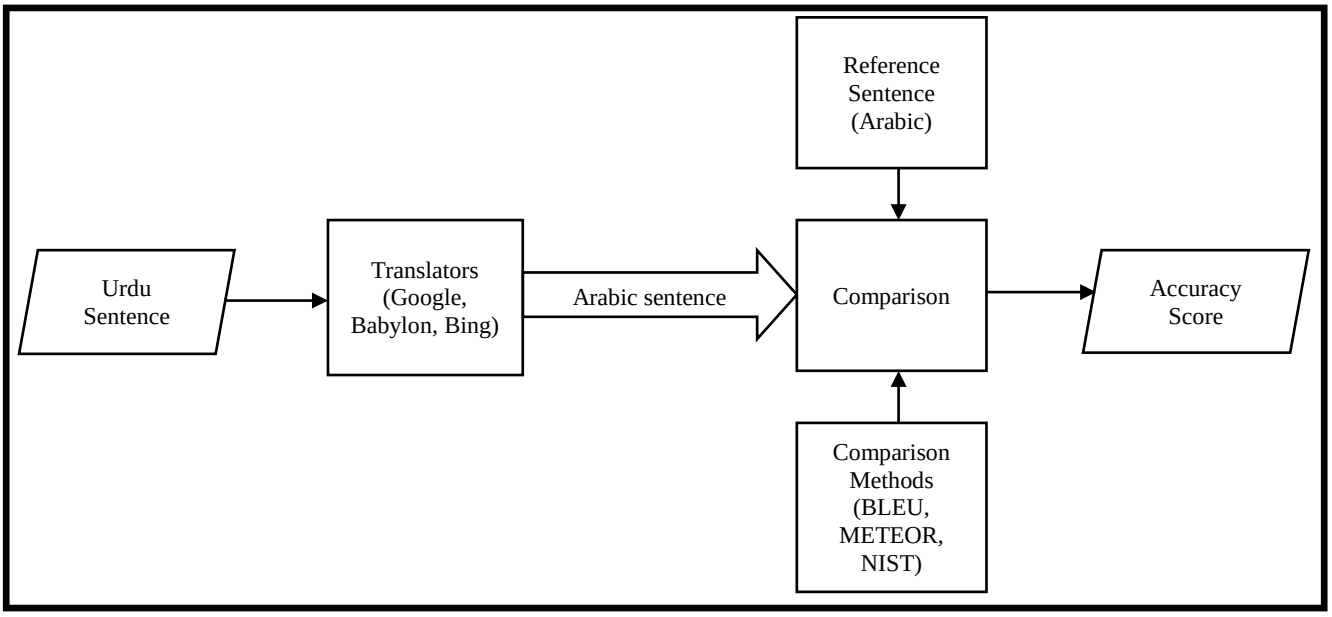


Fig. 1. Block diagram of our comparative evaluation.

TABLE II. EXAMPLE OF ALL TYPE OF SENTENCES USED IN CORPUS

Categories/ Sentence Type	Urdu Source Sentences	Arabic Reference Sentences
Declarative	ملازمین نے ایک لمبی چھٹی لے لی	اخذ الموظفون إجازة طويلة
Exclamatory	کاش وہ وقت پر آجائے!	ليته يحضر على المود!
Imperative	اس کرسی پر بیٹھو	اجلس على ذلك الكرسي

A. BLEU

The BLEU score is calculated by comparing each translated sentence and then comparing with the reference sentence. The average of these scores is computed by averaging them with the corpus size to find the translation accuracy. It is noteworthy that the evaluation does not take into consideration the grammar correctness of the translation. BLEU technique is constructed and put in place to calculate the quality at corpus level. The use of BLEU technique to evaluate the quality of individual sentences always gives an output that lies between 0 and 1. These values tell the readers how similar the reference and candidate sentences (translator output) are. Words with values closer to 1 are closer to the reference translation.

In our case, BLEU divides Urdu sentences into various n-gram sizes, for example, unigrams, bigrams, trigrams and tetra-grams. For each of the four gram sizes, the accuracy for various translators such as Bing, Babylon and Google translator is computed. In the end, for every n-gram sizes, we calculate the n-gram scores of the sentence.

The respective steps to calculate the score for all the n-gram sizes are as follows:

1) Find the total number of common words in every candidate and reference sentence.

2) Then divide their sum over the total number of n-grams in the candidate sentence.

To calculate the BLEU-score these are the steps we need to follow:

1) The first step we need to perform is to calculate the Brevity Penalty (BP) which is calculated by choosing the reference sentence that has the more common *n*-grams length, denoted by *r*.

2) The second step is to compute the total length of the candidate translation, denoted by *c*.

3) Lastly, we need to select the Brevity Penalty to be a reduced exponential in (r / c) as shown in (1).

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \quad (1)$$

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

Where, BP = Brevity Penalty; N = Total number of n-gram sizes; w_n = 1/N and p_n = n-gram precision up to N. The final BLEU score can be calculated using (2) and it is based on Brevity Penalty (*BP*) shown in (1).

A higher BLEU score for a machine translation system implies its superiority to other competitors having lower BLEU scores.

B. METEOR

Another machine translation evaluation technique is known as "Metric for Evaluation of Translation with Explicit Ordering" (METEOR). It premises on the harmonic mean of the unigram precision and recall. This technique is different from the one mentioned above in the sense that it works on the segment level while BLEU works on corpus level.

In METEOR algorithm, the first step is to map an alignment between the reference and candidate sentences. This alignment is established according to the unigram technique. Mapping is also considered to be a line between single word of one sentence with the others. Every single word of candidate sentences must map to either zero or one in the reference sentences. If two alignments map on the same word, then we need to consider the one with the fewest one. The final alignment completed by unigram precision (P) is shown in (3):

$$P = \frac{m}{w_t} \quad (3)$$

Where, m = number of common unigrams in candidate translation and reference translation and w_t = number of unigrams in the candidate sentences. After this we compute the unigram recall (R) by (4):

$$R = \frac{m}{w_r} \quad (4)$$

Where, m is same as above and w_r = number of unigrams in the references sentence. We combine precision and recall to calculate harmonic mean as shown in (5):

$$F_{mean} = \frac{10PR}{R + 9P} \quad (5)$$

This technique is only applicable to the unigrams and not for larger segments. To evaluate the n-gram matches, penalty, p as shown in (6) is used to obtain alignment values.

Processing penalty computations and unigrams are combined with one another in possible groups, where these groups are defined as the combination of unigrams. Longer the adjacent mappings between the reference and the candidate sentence, fewer the chunks are. A translation that is similar to the reference translation gives only one chunk. Penalty (p) can be computed by (6).

$$p = 0.5\left(\frac{c}{u_m}\right)3 \quad (6)$$

Where, c = number of unigrams and u_m = number of mapped unigrams. Final METEOR score can be computed as shown in (7).

$$M = F_{mean}(1 - p) \quad (7)$$

The procedure to calculate the METEOR score for the entire corpus is to get the values for P, R and p and then utilize the formula shown in (7).

C. NIST

NIST stands for National Institute of Standards and Technology. Basically, this is a method devised to check the quality of the text. It is similar to the BLEU metric, because it works on n-grams but, at the same time, it is different from BLEU because it does not calculate the brevity penalty. It is similar, to some extent, to METOR as it computes the precision.

We can calculate the score of NIST by using the formula given below in (8).

$$S_{NIST} = \sum_{n=1}^{N}\left\{\frac{\sum_{all\ w_1 \dots w_n\ in\ sys\ output} info\ (w_1 \dots w_n)}{\sum all\ w_1 \dots w_n\ in\ sys\ output}\right\} \times \exp\left\{\beta \log 2\left[\min\left(\frac{L_{sys}}{L_{ref}}, 1\right)\right]\right\} \quad (8)$$

Where, L_{ref} = the average number of words in a reference translation averaged over all reference translations; L_{sys}= the number of words in the translation being scored; β is chosen to make the brevity penalty factor = 0.5; N indicates the maximum n-gram length; and *info* $(w_1 \dots w_n)$ is

$$info\ (w_1 \dots w_n) = \log 2\left(\frac{the\ number\ of\ occurence\ of\ w_1 \dots w_{n-1}}{the\ number\ of\ occurence\ of\ w_1 \dots w_n}\right) \quad (9)$$

VI. Evaluation Through Example

Here we take an Urdu sentence as an example and its reference translation and machine translations from each MT system.

Source sentence: "رشتہ دار دادا کے گھر میں ہیں"

Reference translation: "الأقارب في بيت الجد"

Arabic Translation of MT system is as follows:

- Google: "الأقارب في البيت القديم"
- Bing: "هم الأقارب في بيت آبائنا"
- Babylon: "الاقارب فى منزل جده"

In this example, we see that all MT systems translate the meaning of "رشتہ دار" ("الأقارب") correctly. For "گھر" Google and Bing matches the reference translation (بيت) while Babylon result is different. All translators give the translation of "دادا" which is not matched with the reference translation.

By calculating the results of above example BLEU score for Google is 0.75, Bing 0.6 and Babylon 0.5. According to METEOR, Google gives 0.75, Bing 0.73 and Babylon 0.5 score. For NIST Google score 2.7, Bing 2.1 and Babylon 1.22. In all the measures, Google is the best in translation as compared to Babylon and Bing.

VII. Results and Discussion

In this section, we repot the results which are generated by our evaluation metrics (BLEU, METEOR and NIST) for the corpus which we mentioned above. We compare the accuracy of each MT system according to three evaluation metrics under separate headings.

A. Comparison of MT Systems Using BLEU Metric

In this section, we exploit BLEU score to compare the performance of each translator. TABLE III shows that by applying the BLEU technique on different types of sentences, Google translator gives 0.1675 score, Babylon 0.0645 and Bing 0.1339 BLEU score for declarative sentences. Google performance is better among all the other translators. For exclamatory sentences, Google gives 0.0577, Babylon 0.0315 and Bing 0.0426 BLEU score. For imperative sentences, Google gives 0.1242, Babylon 0.0459 and Bing 0.0586 BLUE score. By calculating the average of all three sentence types, we see that Google gives 0.1164, Babylon 0.0473 and Bing 0.0783 BLEU score. Average values show that Google's performance is more accurate as compared to those of other translators'.

The average results are also shown in Fig. 2. We can easily see that Google outperforms Bing and Babylon. Google translator, as per BLEU evaluation measure outperforms 28.55% better than Babylon and 15.74% then Bing.

TABLE III. BLEU Score of Each Machine Translator

Translator type	Declarative Sentence	Exclamation Sentence	Imperative Sentence	Average
Google MT System	**0.1675**	**0.0577**	**0.1242**	**0.1164**
Babylon MT System	0.0645	0.0315	0.0459	0.0473
Bing MT System	0.1339	0.0426	0.0586	0.0783

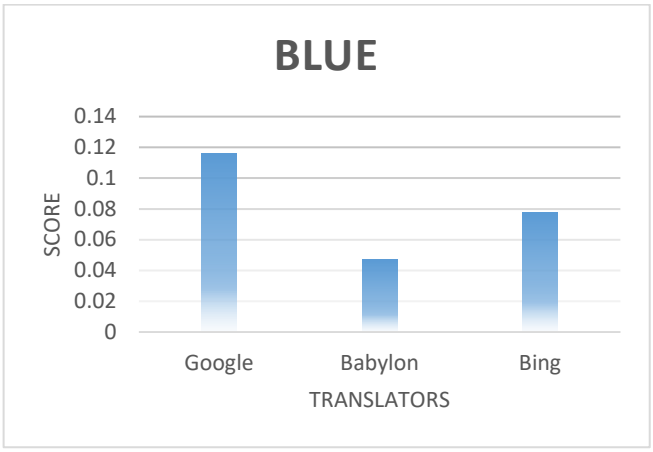


Fig. 2. Comparison of translators by using BLEU metric.

B. Comparison of MT Systems using METEOR Metric

In this section, we exploit METEOR score to compare the performance of each translator. TABLE IV shows that by applying the METEOR technique on different types of sentences, Google translator gives 0.21, Babylon 0.1118 and Bing 0.2014 METEOR score for declarative sentences. METEOR scores for Google and Bing are close to each other and better than that of Babylon'. For exclamatory sentences, Google translator gives 0.16, Babylon 0.14 and Bing 0.16 METEOR score. In the case of exclamatory sentences, Google's and Bing's results are exactly same, and Babylon's results are also very near to them. For imperative sentences, Google gives 0.1558, Babylon 0.0871 and Bing 0.1337 METEOR score. Performance of Bing in this type of sentences is near to Google's but Babylon shows poor performance. Averaging the above results, we see that Google's performance is more accurate as compared to the performance of other translators.

TABLE IV. METEOR Score of Online Machines

Type/ Translator	Declarative Sentence	Exclamation Sentence	Imperative Sentence	Average
Google MT System	**0.2100**	**0.1685**	**0.1558**	**0.1747**
Babylon MT System	0.1118	0.1412	0.0871	0.1130
Bing MT System	0.2014	0.1653	0.1337	0.1600

The average results of TABLE IV are also shown in Fig. 3. Google translator, as per METEOR evaluation measure outperforms 13.74% better than Babylon and 3.28% than Bing.

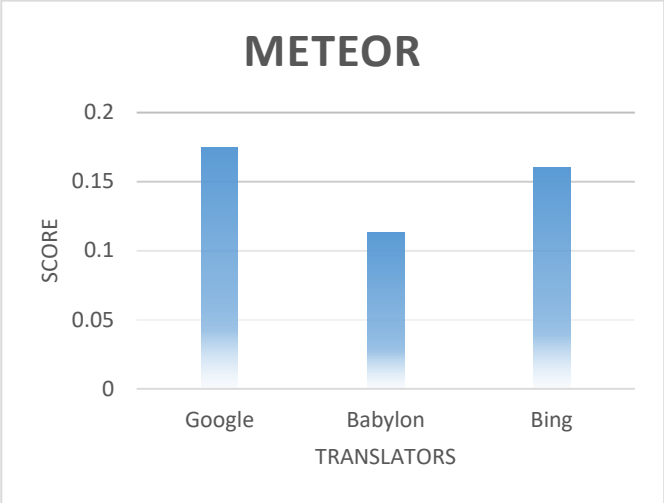
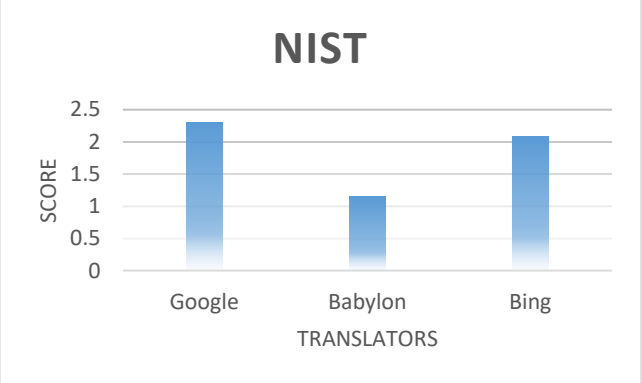


Fig. 3. Accuracy of all online machines using METEOR Technique.

C. Comparison of MT Systems using NIST Metric

In this section, we exploit NIST score to compare the performance of each translator. TABLE V shows performance of MT systems for different types of sentences using NIST metric. TABLE V shows that Google gives 3.14, Babylon 2.0234 and Bing 2.9629 NIST score for declarative sentences. For exclamatory sentences, Google gives 1.5199, Babylon 1.1885 and Bing 1.3881 NIST score. In the case of such sentences, Google and Bing are nearly equal to each other and both are better than Babylon. For imperative sentences, Google gives 2.2591, Babylon 1.3469 and Bing 1.9808 NIST score. Performance of Google in imperative sentences is much better than that of Bing and Babylon. By calculating the average of all sentence types, we see that Google gives 2.306, Babylon 1.151 and Bing 2.089 NIST score. According to this average, Google is the best in accuracy.

The average results of TABLE V are also shown in



. We can see that Google outperforms Bing and Babylon. Google translator, as per NIST evaluation metric, outperforms Babylon by 20.83% and Bing by 3.91%.

TABLE V. NIST SCORE OF ALL ONLINE MACHINES

Type/ Translator	Declarative Sentence	Exclamation Sentence	Imperative Sentence	Average
Google MT System	**3.1489**	**1.5199**	**2.2591**	**2.3060**
Babylon MT System	2.0234	1.1885	1.3469	1.1510
Bing MT System	2.9629	1.3881	1.9808	2.0890

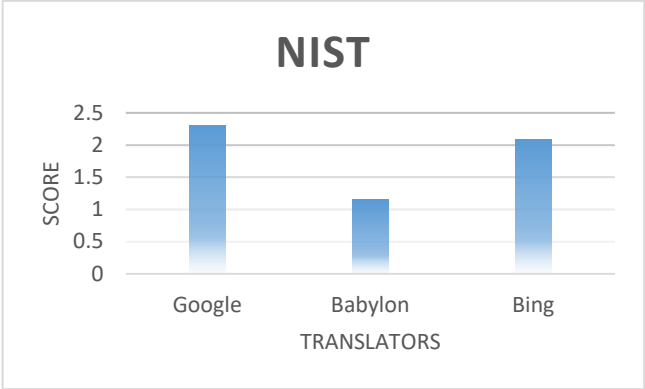


Fig. 4. Accuracy of all online machines using NIST Technique.

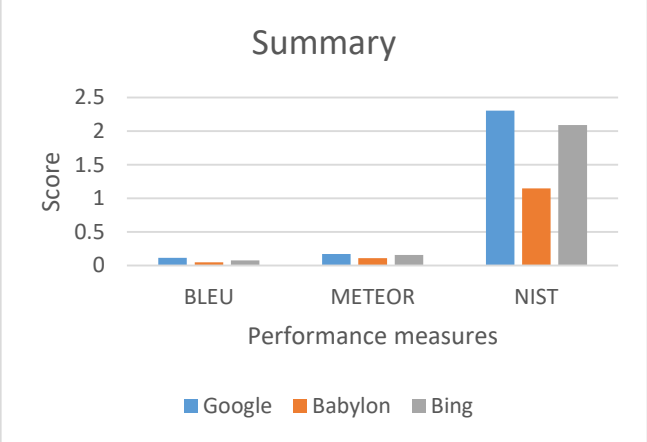


Fig. 5. Summary of tools with respect to evaluation measures.

Comparing results in all techniques BLEU, METEOR and NIST, it is concluded that Google always outperforms Babylon and Bing translators. Fig. 5 shows the summary of all results of all translators w.r.t BLEU, METEOR and NIST metric.

VIII. SUMMARY AND FUTURE WORK

In this paper, we compare three machine translators (Google, Bing and Babylon) for translating Urdu sentences to Arabic sentences by using three performance evaluation metrics (BLEU, METEOR and NIST). The corpus used in this research contains three different types of 159 Urdu sentences and their respective Arabic sentences. Our results show that Google translator, on the average, outperforms Bing and Babylon by 15.74% and 28.55% in BLEU technique, 13.74% and 3.28% in METEOR technique, 20.83% and 3.91% in NIST technique respectively. This study is helpful for those who want to use online machine translators for Urdu to Arabic translation.

We will develop our own Urdu to Arabic machine translation system by exploiting hybrid technique comprising template based and rule based approach. We expect to have

better results than the available online machine translators. In future, we will also build a large corpus for evaluation MT systems.

REFERENCES

[1] Durrani, Urdu Informatics, Center of Excellence for Urdu Informatics, National Language Authority, Islamabad, Pakistan, 2008.

[2] C. Holes, Modern Arabic: Structures, Functions, and Varieties, Washington, D.C.: Hopkins Fulfillment Services, 2004.

[3] D. Kenny, Lexis and Creativity in Translation: A corpus based approach, Routledge; New Ed edition (1 Jan. 2001), 2014.

[4] R. Harshawardhan, Thesis on "Rule based machine translation system for English to Malayalam language," 2011.

[5] A. Gehlot, V. Sharma, S. singh and A. Kumar, "Hindi to English transfer based machine translation system," International Journal of Advanced Computer Research, vol. 5, no. 9, 2015.

[6] M. Osborne, C. Bruch, Chris and D. Talbot, "Statistical machine translation with word-and sentence-aligned parallel corpora," in The 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[7] A. Sameh, "Interlingua-based machine translation systems: UNL versus Other Interlinguas," in 11th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt, 2011.

[8] S. M. Kadhem and Y. R. Nasir, "English to Arabic example-based machine translation," IJCCCE, vol. 15, no. 3, pp. 47-63, 2015.

[9] C. Vertan, "Knowledge based machine," 2005. [Online]. Available: https://nats-www.informatik.uni-hamburg.de/pub/User/IntensiveCourseInMachineTranslation/kbmt.pdf.

[10] S. Hunsicker, C. Yu and C. federmann, "Machine learning for hybrid machine translation," Seventh Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 312-316, 7,8 June 2012.

[11] 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, vol. 5, no. 11, pp. 68-73, 2014.

[12] M. D. Okpor , "Machine translation approaches: Issues and challenges," IJCSI International Journal of Computer Science, vol. 2, no. 5, p. 159, 2014.

[13] X. Song, T. Cohn and L. Specia, "BLEU deconstructed: Desiging a better MT evaluation metric," International Jornal of Computational Linguistics and Applications, vol. 4, no. 2, pp. 29-44, 2013.

[14] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in In Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014.

[15] P. Thi Le Thuyen and V. Trung Hung, "Results Comparison of machine translation by Direct translation and by Through intermediate language," International Journal of Advance Research in Computer Science and Management Studies, vol. 3, no. 5, pp. 15-20, 2015.

[16] C. Kit and T. M. Wong, "Comparative evaluation of online machine translation systems with legal texts," Law Libre. J., no. 100, pp. 291-321, 2008.

[17] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa and M. Izzat, "Evaluating English to Arabic machine translation using BLEU," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no. 1, pp. 66-73, 2013.

[18] M. Aiken, K. Ghosh, J. Wee and M. Vanjani, "An evaluation of the accuracy of online translation systems," Communications of the IIMA, vol. 9, no. 4, pp. 67-84, 2009.

[19] M. W. Aiken and Z. Wong, "Spanish-to-English translation using the Web," in Southwestern Decision Sciences Institute, Oklahoma City, Oklahoma, March 9 – March 13, 2006.

[20] M. Vanjani, M. Aiken and J. H. Ablanedo Rosas, "Efficacy of English to Spanish automatic translation," International Journal of Information and Operations Management Education, vol. 2, no. 2, pp. 194-210, 2007.

[21] 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, March 2014.

[22] Q. Ying, Q. Wen and J. Wang, "Automatic evaluation of translation quality using expanded N-gram co-occurrence," in Natural Language Processing and Knowledge Engineering (NLP-KE). International Conference on, IEEE, 2009.

[23] A. Lavie, K. Sagae and S. Jayaraman, "The significance of Recall in automatic metrics for MT evaluation," in AMTA, Washington DC, 2004.