Evaluating Machine Translations from Arabic into English and Vice Versa

Riyad Al-Shalabi, Ghassan Kanaan
Amman Arab University
Amman, Jordan
shalabi@aau.edu.jo, gkanaan@aau.edu

Huda Al-Sarhan, Alaa Drabsh & Islam Al-Husban
Jordan University of Science and Technology
Irbid, Jordan
johnibrahim12@cit.just.edu.jo, afdarabseh12@cit.just.edu.jo, islam_husban@hotmail.com

Abstract—**Machine translation (MT) allows direct communication between two persons without the need for the third party or via dictionary in your pocket, which could bring significant and per formative improvement. Since most traditional translational way is a word-sensitive, it is very important to consider the word order in addition to word selection in the evaluation of any machine translation. To evaluate the MT performance, it is necessary to dynamically observe the translation in the machine translator tool according to word order, and word selection and furthermore the sentence length. However, applying a good evaluation with respect to all previous points is a very challenging issue. In this paper, we first summarize various approaches to evaluate machine translation. We propose a practical solution by selecting an appropriate powerful tool called iBLEU to evaluate the accuracy degree of famous MT tools (i.e. Google, Bing, Systranet and Babylon). Based on the solution structure, we further discuss the performance order for these tools in both directions Arabic to English and English to Arabic. After extensive testing, we can decide that any direction gives more accurate results in translation based on the selected machine translations MTs. Finally, we proved the choosing of Google as best system performance and Systranet as the worst one.**

Index Terms: **Machine Translation, MTs, Evaluation for Machine Translation, Google, Bing, Systranet and Babylon, Machine Translation tools, BLEU, iBLEU.**

I. INTRODUCTION

With increasing number of languages around the world, the need to find a reliable and efficient intermediate layer to communicate becomes necessary. Linguists solve this problem, in many ways, such creating dictionary or even using hiring someone to do the translation as third party even thus. These solutions are expensive and take longer time.

Translation is the process of transferring the meaning of a source-language text to its equivalent target-language. If you want to translate an Arabic sentence to English manually minimally, we need:

A dictionary mapping each Arabic word to English word(s).
Rules to represent the Arabic sentence structure.
Rules to represent the English sentence structure.
Rules to relate these two structures together.

A method for selecting one alternative translation over another one. Therefore, with development of technology, Machine translators MTs come to solve these issues. Machine translation (MT) is the process of automatic translation from one natural language to another using computers, and it helps us to better understand the meaning. The Translator (Machine Translation tool) must interpret and analyze all the features of the text, this process requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language. Since the work of MT systems considers the automatic process, so that it surely saves time, effort and not expensive.

How do we measure translation performance? As Papineni, Kishore and et al. says the closer machine translation output to a professional human translation is the best [1].

Hence, the challenging part is that a single source sentence could be translated to multiple target sentences. These translations may vary in word choice or in word order even when they use the same words. This problem may be solved in many approaches such as rule-based, statistical-based, example –based and hybrid.

Rule Based Machine Translation (RBMT) which also known as Knowledge-Based Machine Translation is an approach of translation based on linguistic information about the source and target like word meaning and grammar rules retrieved from bilingual dictionary and bilingual grammar [2].

Statistical Machine Translation (SMT) is a paradigm where generated translating based on statistical models whose parameters derived from the analysis of bilingual text corpora [2].

Moreover, there are different MTs available to users either offline or online so which tool should we choose. Evaluating MT by its output considers a good way to make the decision. MT evaluation: is the process focusing on evaluating the output of machine translation, rather than focusing on performance or usability evaluation. There are two evaluation approaches:

Human evaluation: Automatic Language Processing Advisory Committee, Advanced Research Projects Agency.

Automatic evaluation matrices: BLEU, NIST, Word error rate, METEOR.

Considering the Benefits of automatic evaluation metrics: they are Fast, Low cost, and automatically optimize system performance gives us an insight that is could be the best choice.

IRJECE

In this paper, we use the iBLEU metric that is an enhanced version of the BLEU metric which is an automatic evaluation tool based on the closeness of the candidate translation (MT output) to the reference translation.

The rest of this paper is structured as follows. We will consider Methodology in section II. Section III will discuss project experiments and results. After that, section IV will analyze the results of the experiments. Finally, we conclude and discuss some futuristic works in section V.

II. METHODOLOGY

With the increased number of machine translations (MTs) the need to decide the best one becomes necessary. In this paper we tried to help the users to make the right decision. In this section we define all necessary points in the paper.

A. Project Scope

MT system is an automated way for translating one language into another. There are a lot of languages around the world and most of machine translators support bilingual translation between these languages. In our paper, we select just two languages to do all the experiments; Arabic and English language. Since Arabic language considered to be one of the most familiar national languages and it is the language which we decide to evaluate how the MTs deal with it. In addition English language is selected because it is the most famous language among people around the world. We evaluate how the MTs translate the sentence from Arabic to English and vice versa.

B. Data Set

Since all MT translate sentence by sentence we select 60 sentences to complete the experiments. The sentences are selected from journals, articles, papers which spread over the internet. All the selected sentences have been posted in both Arabic and English language. We take the translation as a reference to compare the outputs of MTs with the correct translation.

C. Machine Translation Tools

Many MTs are available either online of offline. All of them have its characteristics and differ in some points like correctness and efficiency. In the paper experiments, we select the most famous MTs that are used by people through the internet; such as Google [5], Babylon [6], Microsoft Bing [7] and Systranet [8] MTs.

D. Evaluation Tool

As mentioned previously, there are two approaches to evaluate the MTs; Human and Automatic evaluation. In our experiments we select the one which is the closest to human estimation, which is upgraded version of BLEU tool.The main idea for iBLEU is to measure the closeness between MT system translation output and the reference which has the correct translation obtained from linguistic people.

E. System Algorithm

We set and follow our procedure to achieve our contribution. The main steps are described in details in the algorithm shown in Fig.1.

Algorithm: Evaluating MT systems by iBLEU metric.
Input: A set of sentences
Output: iBlEU scores for each sentence translated by MT system.

BEGIN

a. Initialization Sentences & Project Folder

1. Create four folders inside a head NLP project folder (Dataset, Results, and XmlSheet)
2. Extract all sentences from their resources. And create a world file (Dataset.txt) which contains the dataset.
3. Assign an ID for each sentence. The range of IDs from 1 to 60.
4. Insert the sentences into a Table.

ID	Sentence in Arabic	Sentence in English

b. Translate the Sentences

1. Create four folders for MT systems' output (Google_res, Babylon_res, Bing_res and Systrainet_res) into dataset folder.
2. Inside each one, we create two word documents (MT_output_from_Arabic to English.txt & MT_output_ from English to Arabic.txt).
3. Create a Table inside MT_output_from_Arabic to English.txt by the following header.

ID	Sentence in Arabic	Reference	MT translation output

4. Create a Table inside MT_output_ from English to Arabic.txt by the following header.

ID	Sentence in Arabic	Reference	MT translation output

5. Translate each sentence by each MT and write the output into the tables in the documents.

c. Write the Xml Sheet Files

1. Inside XmlSheet folder we create four folders for each MT.
2. Inside each folder we create two sub-folders also to store the xml files (Arabic to English) & (English to Arabic).
3. Inside each sub-folder, create three xml sheet files(MT_out.xml, Reference.xml, and src.xml files).

d. Evaluate each MT By Ibleu Metric

IRJECE

1. Run the iBLEU tool 8 times for each sub folder.
2. Upload the three files into iBLEU. As shown in Figure 2.
3. Create four Excel files inside results folder.
4. Store the iBLEU score for each sentence.

ID	Score

END

Fig 1. Pseudo Code of the Algorithm

III. EXPERIMENTS AND RESULTS

A. Results

In this section we will present our experiments after applying our algorithm. Figure 2 shows how the files were uploaded to iBLEU tool, and it shows a snapshot from iBLEU after finding the scores.

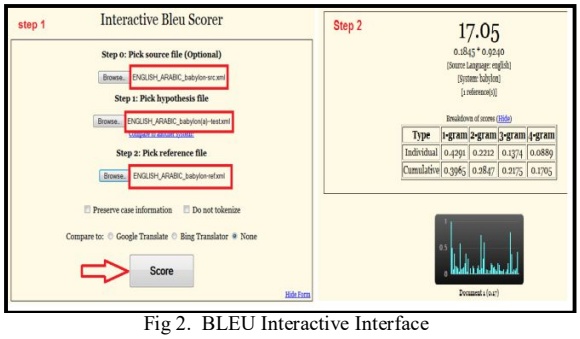

Fig 2. BLEU Interactive Interface

In Fig. 2 we note that the blue as an example, gets two types of scores; individual and Cumulate for single and all sentences, respectively to translate sentences from English to Arabic by Babylon MT.Fig.3 shows how the iBLEU presents the scores for good and bad transaction.



Fig 3. Ibleu Examples Snap Shots

As we see the sentence with ID equal 1 has iBLEU score '1' which means a good MT output. While the sentence with ID equal 17 has iBLEU score '0.02' which indicate bad MT output.

After finding all ability scores for all MT systems, we combine them as shown in Fig. 4 and Fig. 5.

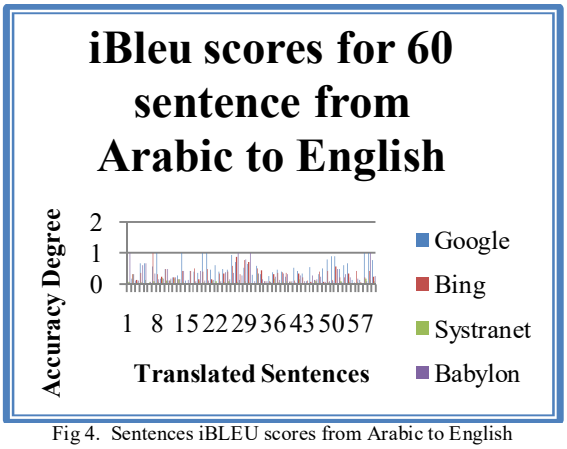

Fig 4. Sentences iBLEU scores from Arabic to English

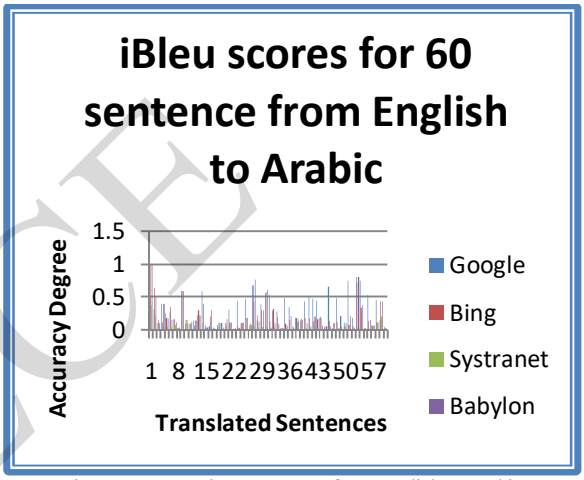

Fig 5. Sentences iBLEU scores from English to Arabic

As we note, the results are ambiguous, and we need a way to normalize them, so we convert the results to be more formal as we will show in the next section. In order to test how the MT systems deal with sentences size, we divide all sentences into three categories. And we see how the MT are affected by the sentence length.

Short sentences → 2-10 words.
Medium sentences → 11- 20 words.
Long sentences → 21-35 words.

B. Examples

In this section, we exhibit two examples in order to explain the idea in the details. We study how the sentence translated by the four MT systems, and check the iBLEU scores for each one. The first example presents a short sentence translated from English to Arabic. While thesecond sentence considers a long sentence which was translated from Arabic to English.

Example1:

	Source sentence: Praises be to Allah	iBLEU Scores
Google	الحمد لله رب العالمين	0.32
Babylon	بحمد الله	0.50
Bing	أن الحمد لله	0.64
Systranet	تمجيد إلى الله	0.23
	Reference: الحمد لله	

Example 2:

	Source sentence: يا بني أريدك أن تدق مسماراً في سياج حديقتنا الخشبي في كل مرة تحتاج لتوجيه غضبك ضد شيء وفقدت أعصابك	iBLEU Scores
Google	I want you to hammer a nail into our garden fence every time you need to direct your anger against something and you lose your temper	0.89
Babylon	Son, I want you to ring bolt in wooden garden fence in each time you need to direct your anger against something and lost chill out	0.49
Bing	My son wanted you to hammer a nail at wooden garden fence every time you need to direct your anger against something and lost your temper	0.58
Systranet	Oh brown wanted you that nail in our wooden fence of garden knocks every time orientation of anger needs for you against thing and your nerves lost.	0.04
	Reference: My son, I want you to hammer a nail into our garden fence every time you need to direct your anger against something and you lose your temper	

IV. EXPERIMENTAL ANALYSIS

In order to test the effectiveness of the Machine translation systems, the results of translating 60 sentences obtained by four selected MTs - Google, Bing, Systranet and Babylon - are compared with the original translation that conducted by linguistic people. Then check the accuracy of translation in both directions Arabic to English and English to Arabic using iBLEU tool.

In the first case, we will evaluate each sentence regardless of its length or counting number of words in each. Then, compute the average of accuracy when translating from Arabic to English and vice versa.

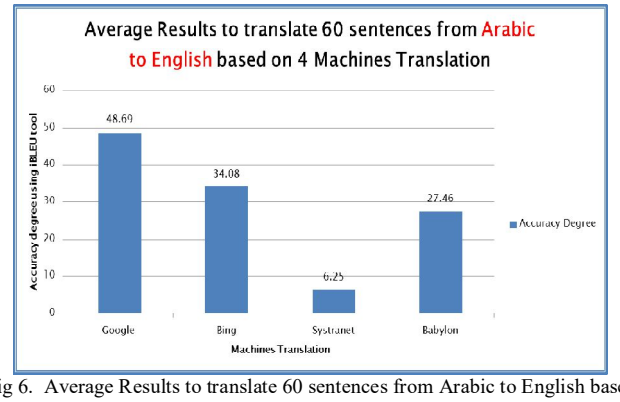

Fig 6. Average Results to translate 60 sentences from Arabic to English based on four Machine Translations by using iBLEU tool

Figure 6 shows the accuracy rate that computed by iBLEU tool when translating sentences from Arabic to English. Based on the experiences, the results prove that the Google Translation outputs more accurate than others do. This indicates to consider the Google Translator as more effective and efficient than other systems. In contrast, the worst machine translation is Systranet, since it provides a translation less compatible with the reference version. Therefore, if the direction of translation from Arabic to English, the sequence of MTs could be arranged in descending order from best to worst as follows:

Google → Bing → Babylon → Systranet

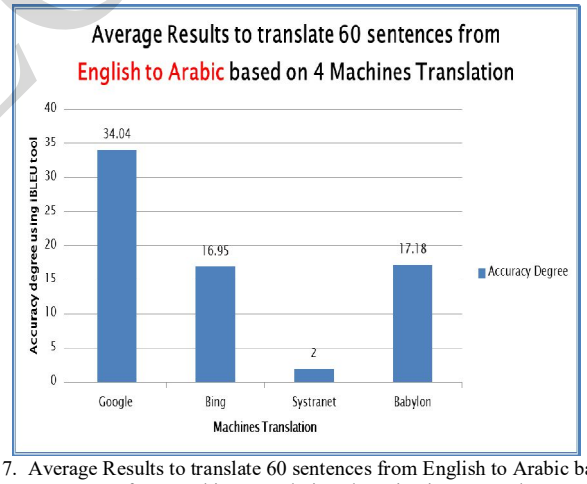

Fig 7. Average Results to translate 60 sentences from English to Arabic based on four Machine Translations by using iBLEU tool

We can notice from Fig. 7 which deals with translating the sentences from English to Arabic the superiority of Google Translator again over other machine translations. It gives the highest accurate degree and provides 34.69 referring to the proximity of the original translation. The second machine tool in efficiency is Babylon translator with result 17.18. Third tool that gives a slight difference in accuracy is Bing translator. Finally, the worst tool in terms of the differences about the reference

IRJECE

copy is Systranet tool. Thus, the descending order for the MTs in the direction from English to Arabic is:

Google → Babylon → Bing→ Systranet

The TABLE I below summarizes the previous results in both directions: Arabic to English (A – E) and English to Arabic (E – A) using iBLEU tool.

Table I. Accuracy degree for four Machine Translations in both directions Arabic to English (A – E) and English to Arabic (E – A) using iBLEU tool

Direction	Machine Translation Type			
	Google	Bing	Systranet	Babylon
A – E	48.69	34.08	6.25	27.46
E – A	34.04	16.95	2	17.18

Table I values show that the accuracy degree of translation from Arabic to English is higher than from English to Arabic in all cases. To illustrate the difference, Fig. 3 below represents the values shown in the previous table.

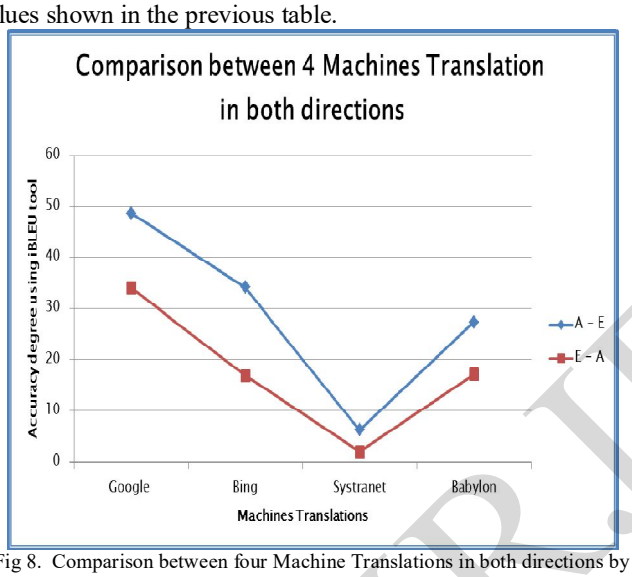

Fig 8. Comparison between four Machine Translations in both directions by using iBLEU tool

Fig. 8 shows that all the MTs work better when the direction is from Arabic to English. In other words, Machine translations could give satisfying results because they are closer to human translation.

In the second case, we classify the original dataset into three categories based on sentence length. This can apply through counting the number of words each sentence by using an online tool [9]. Then, three classifications will produce: Short Sentences, Medium Sentences and Long Sentences. The Table Ii displays the values of three sentence classifications according to four machine translations in direction from Arabic to English.

Table II. Comparison Between three classifications of the sentences According to four machine translations (MTs) in direction from Arabic from Arabic to English

Arabic to English				
Sentences Classification	Google	Bing	Systranet	Babylon
Short Sentences	0.516	0.336	0.116	0.383
Medium Sentences	0.427	0.232	0.061	0.16
Long Sentences	0.484	0.383	0.086	0.311

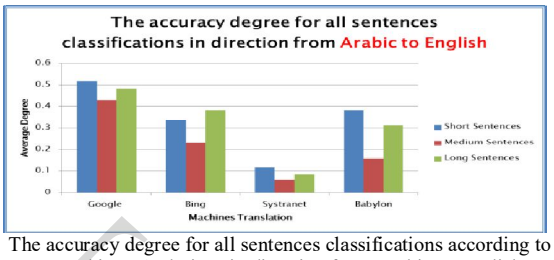

Fig 9. The accuracy degree for all sentences classifications according to four machine translations in direction from Arabic to English

Fig. 9 illustrates the continuity of Google translator to take first rank in terms of the accuracy rate, whilst Systranet translator has the last rank. In addition, according to the short sentences, the Babylon translator has the second rank and Bing translator gains a third rank. On the other hand, Bing translator outperforms Babylon translator and gets the second rank for both classifications belong to medium and long sentences. We can summarize the sequence of MTs in descending order as follows:

- Short sentences

Google → Babylon → Bing → Systranet

- Medium sentences

Google → Bing → Babylon → Systranet

- Long sentence

Google → Bing → Babylon → Systranet

Table III compares between three classifications according to four machine translations in direction from English to Arabic and Fig. 10 clarifies this issues.

Table III. Comparison Between three classifications of the sentences According to four machine translations (MTs)in direction from Arabic from English to Arabic

English to Arabic				
Sentences Classification	Google	Bing	Systranet	Babylon
Short Sentences	0.219	0.232	0.097	0.228
Medium Sentences	0.316	0.142	0.034	0.153
Long Sentences	0.373	0.208	0.048	0.180

IRJECE

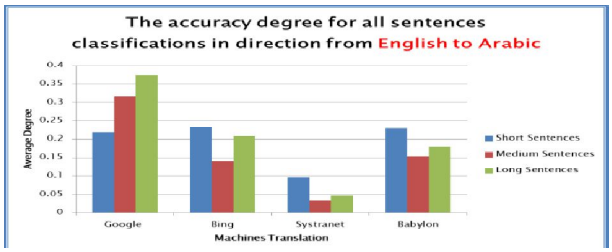

Fig 10. The accuracy degree for all sentences classifications according to four machine translations in direction from English to Arabic

In Fig. 10, the sentences are translated from English to Arabic. The results show the decline of the Google translator by taking the third rank in precision after Bing and Babylon translators in short sentences. While Google occupies the first degree in the medium and long categories, and this demonstrates the strength and efficiency of the Google MT when the length of sentences increases. Systranet translator remains the worst one in all cases and in both directions.In addition, there is variation in the results of the Bing and Babylon translators according to three classifications. Since Bing translator is the best in the short sentences, second in the long sentences and third in the Medium sentences. Whereas Babylon translator gets the second rank in short and medium sentences and third rank in the long sentences. The summary of the results as follows:

- Short sentences

Bing → Babylon→ Google → Systranet

- Medium sentences

Google→ Babylon → Bing → Systranet

- Large sentences

Google → Bing → Babylon→ Systranet

V. CONCLUSION AND FUTURE WORK

This paper presents a powerful automatic evaluation tool called iBLEU that provides scoring information in a visual, fast, and easy to understand manner. We have compiled our dataset from different articles, forums and political websites. By getting advices from linguists, the sentences are translated in Arabic and English. After that, we compare the reference copy with four famous machine translations (Google, Bing, Babylon and Systranet). In this way we achieve our goal of finding the convergence or closeness between machine translation and human translation to determine the percentage of accuracy for each sentence in both directions whether Arabic to English or English to Arabic.

From the obtained results, we can conclude that Google Translate outputs are slightly better than the outputs of Bing, Babylon and Systranet MT Systems in both directions, whilst the Systranet was the worst at all levels. Moreover, we can list the sequence of MTs from Arabic to English in descending order from the best to the worst as follows: (1) Google (2) Bing (3) Babylon (4) Systranet. Furthermore, the sequence of MTs from English to Arabic is (1) Google (2) Babylon (3) Bing (4) Systranet. In addition, the results prove that all MTs work better from Arabic to English.

We conduct extra experiments after taking length statement criterion into consideration. Therefore, the original dataset are divided into three classifications: short, medium and long sentences. The measurements indicate that Google machine translation is slightly better than other MT Systems related to three categories from Arabic to English. However, when translate from English into Arabic, Google MT still the best in the all categories except the first one referring to short sentences. Thus the Bing and Babylon translators have achieved better results than Google MT system.

For future work, we plan to expand our dataset to include more than 100 sentences. Then, repeat the previous experiments whether taking consideration the length of sentence or ignoring this criterion. Thereafter, we check the accuracy degree based on selected machine translations. Another goal aim to be achieved is dividing the collected sentences into 4 types: exclamatory, declarative, interrogative, and imperative respectively. Each type should contain an equal number of sentences and measure the accuracy of four machine translators using iBLEU metric. Final goal we hope to achieve is to find new powerful automatic evaluation tools rather than iBLEU, then compare the new results with the former one and find the relation if possible.

REFERENCES

[1] P., Kishore, and et al, "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
[2] W. John Hutchins and H. Somers, “An introduction to machine translation.”. London : Academic Press, 1992.
[3] T. Hailat, T., M. Al-Kabi, M.,I. Alsmadi, E. Al-Shawakfa., "Evaluating English to Arabic machine translators," Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on , vol., no., pp.1,6, 3-5 2013.
[4] M. Alawneh, T. Sembok,and M. Mohamad, "Grammar-based and example-based techniques in machine translation from English to Arabic," Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on , vol., no., pp.1,6, 2013.
[5] A. Almahairi, A. Almahairi, K. Cho, N. Habash, and A. Courville. First Result on Arabic Neural Machine Translation. arXiv preprint arXiv:1606.02680. 2016.
[6] https://translate.google.com/ Accessed June 1, 2014.
[7] http://www.babylon.com/ Accessed June 1, 2014.
[8] http://www.bing.com/translator/ Accessed June 1, 2014.
[9] http://www.systranet.com/translate/ Accessed June 1, 2014.
[10] http://textmechanic.com/Count-Text.html Accessed April 22, 2014.

IRJECE