# Assessment of Machine Translation Output: A comparative study between Human and Automatic Models

**Article** *in* Cihan University-Erbil Scientific Journal · January 2018

DOI: 10.24086/cuesj.si.2018.n1a9



**1 author:**

Fereydoon Rasouli
Cihan University
**7** PUBLICATIONS **9** CITATIONS

SEE PROFILE

# Assessment of Machine Translation Output: A comparative study between Human and Automatic Models

Fereydoon Rasouli
Translation Department, Cihan University-Erbil
Fereydoon.rasouli@cihanuniversity.edu.iq

## Abstract

In this study, it is attempted to make a comparison between two common methods of evaluation of machine translation (MT) output (Human and Automatic MT evaluation). Materials of the study have been selected from economical texts. Twenty English sentences and their Persian translation were selected from "translating of economic texts" book published by Payam-e-Nour University. To assess translated sentences humanly 20 Ma students of translation studies participated in this study as evaluators. In order to evaluate sentences automatically, BLEU method of Mt output evaluation was applied. According to the findings of the study both methods of evaluation lead to the same results, however , human evaluation method is more precious than automatic evaluation methods, at the same time automatic evaluation methods is faster and more time saving than human evaluation methods.

**Keywords:** Automatic Evaluation Methods, Human Evaluation Methods, Machine Translation.

## Introduction

The evaluation of natural languages processing systems, including |MT outputs are very important, especially to the users of Mt systems, system developers or researchers (Doug Arnold & Louisa Sadler, 1993). As Arnold and Sadler (1993) stated users of MT systems are interested in quality and cost-effectiveness of translation, for developers answering to this question is more important that whether their efforts making their system better, and finally, from researchers point of view the evolution of systems

DOI: 10.24086/cuesj.si.2018.n1a9

that embody theoretical ideas provides a partial evaluation of those ideas, and, by indicating deficiencies, may provide hints about research priorities and areas of research.

In 1966, the ALPAC report funded and sponsored by US government was published advising against further investment in MT. the report concluded that machine translation is slower, less accurate and more expensive than human translation (ALPAC, 1966).

In 1992, the AMTA workshop devoted to MT evaluation was held provided the basis for future directions. In this workshop JEIDA report was presented by Japan's Electronic Industry Development Association entitled Methodology and Criteria on MT evaluation. This report stressed the importance of judging system according to the context of use and user requirements. in1993, the machine translation journal came to existence devoted to MT evaluation. Between 1992 and 1994, DARPA (defense advanced research project agency) was also working seriously on MT evaluations.

Between 1992 and 1999 EAGLES (Expert Advisory Group on Language Engineering) set up by EC had several aims one of which was to propose standards, guidelines and recommendation for good practice in evaluation of language engineering products. Most of evaluation of MT takes place under contract and often under confidentiality agreements. Consequently there is little constructive criticism of methodology. A major deficiency is that many evaluation are undertaken by people with little or no expertise in MT techniques, unable to judge what is possible and what is unrealistic, unable to estimate the potential rather than current performance, on the other hand, evaluations made by MT researchers are often minimal and misleading: the demonstration of a system with a carefully selected set of sentences or sentence types is not the basis for claims about a large scale system (Saedi, C at al. 2009).

As M. Dobrinkat (2008) believes the main purposes of evaluation of MT systems are:

- allows the comparison of different MT systems or different versions of one system. Evaluation helps to determine which system is the best in a certain aspect or for some specific purpose or domain.
- allows optimization of performance by finding system modifications that yield improved evaluation results.

Gerber (2001) believes that in order to measure the quality of MT we should be able to measure the content of text. He argues that text characterization should include the real-world state of affairs as well as communicative goal of a piece of text.

Van Slype (1979) classified evaluation of MT into two subcategories: macro and micro evaluation aspects. Macro evaluation considerers evaluation aspects with regard to the user requirements such as goodness of translation whereas micro evaluation considers the sources of insufficiency and so tries to look inside the translation system black box (Van Slype, 1979).

## Aspects of MT evaluation

- **Intelligibility** measures the ease with which a translation can be understood (M. Dobrinkat, 2008). As Douglas Arnold et al (1994) noted intelligibility is a traditional way of assessing the quality of translation where it is affected by grammatical errors, mistranslations and untranslated words. Scoring system is assigned to do evaluating of intelligibility that reflect top marks for those sentences that are so badly degraded as to prevent the average translator /evaluator to determine which translation is acceptable in the context (ibid, p.161).

- **Accuracy** because a highly intelligible output sentence is not necessarily a correct translation of the source sentence it is important to check whether the meaning of the source language is preserved in translation the property which assesses this aspect of translation is called accuracy or fidelity (ibid, p. 162). As M. Dobrinkat (2008) argued accuracy measures how much of the information in the source language is successfully transferred to the target language. The evaluation procedures of accuracy, based on Douglas Arnold et al (1994), is similar to the one used for scoring of intelligibility. With this difference that it's highly source text oriented so that they can compare the meaning of input and output sentences.

  Because accuracy scores are often closely related to the intelligibility scores, accuracy scores are much less interesting than intelligibility scores. High intelligibility normally means high accuracy (ibid p. 163).

- **Usability** basically refers to meets users' expectation if an output of MT is usable for those how want to buy and use it (M. Dobrinkat**,** 2008)**.**

## Time- efficiency aspects of MT evaluation

As Dobrinkat (2008) noted reading time, correction time and evaluation time are three aspects of time efficiency of MT evaluation. They respectively refer to time required for reading the translated text, the amount of time that translator should spend for editing and correcting the MT out-put and the required time for a MT system to perform its task.

## Linguistic aspects of MT evaluation

- **Semantic relationships** refer to evaluating of the semantic aspects which translated by MT system correctly.
- **Lexical evaluation** measures the amount of common words between the reference translation and the translation performed by MT system.
- **Syntactic and morphological coherence** measure consistency among syntax and morphology of the MT output.

## Human evaluation

### 1. Language and Usability Perspective

A traditional method of evaluating MT output is to look at output and judge by hand whether it is correct or not (Philipp Koehn, 2010). This task is usually done by bilingual evaluators who understand both the input and output languages are Evaluation is done at sentence level, but a longer document context may be essential to carry out the judgments. (Philipp Koehn, 2010). The quality of MT output can be judged from the language perspective or the usability perspective. According to White (2003a) an intuitive way to do assessment is by rating of intuitive judgment of the goodness of the translation. Evaluators are asked to rate a translation, normally presented sentence by sentence, in terms of their intuitive judgment. A set of attributes such as fidelity, intelligibility adequacy and fluency, determine the goodness of translation. More common approach of human evaluating is to use a graded scale when eliciting judgment from human evaluators. Two common criteria in human evaluation are fluency and adequacy. Phillip Koehn (2010) illustrated that how these criteria are scored:

*Adequacy*

5 all meaning

4 most meaning

3 much meaning

2 little meaning

1 none

**Fluency**

5 Flawless English

4 Good English

3 Non-native English

2 Diffluent English

1 Incomprehensible

**2. Error Analysis**

While quality assessment evaluation appraises the "goodness" of translation, error analysis judges the translation quality from the opposite perspective, that is, by measuring the "badness" of translation. It starts from counting the errors in the translation, which shows "the amount of work required to correct 'raw' MT output to a standard considered acceptable as a translation" (Hutchins and Somers, 1992). An error is defined as "any deviation from TL (target language) intelligibility or translation accuracy" by Schwarzi (1999). He mentions that the error analysis method is more reliable than the quality assessment method, because identifying of errors is relatively more objective and consistent among evaluators than rating the goodness of translation.

## Automatic Evaluation Methods

Through human evaluation different aspects of translation such as adequacy, intelligibility and accuracy are assessed. Performing this comprehensive task is also expensive and time consuming; therefore, it is preferred to have an automatic method for assessing the quality of machine translation output by which make it clear, whether our system got better after a change, or not. This is the objective of automatic machine translation evaluation (Philipp Koehn, 2010).

Different Automatic MT is used to assessing MTs output here some of them which are the most popular methods among MT evaluators and researchers presented briefly:

**Meteor**

Meteor is proposed initially in 2004 and was designed to improve correlation with human judgments of MT quality at the segment level (Lavie et al., 2004). METEOR evaluate a translation by computing a score based on clear word to word matches between the translation(s) and a given reference translation (Abhaya Agarwal & Alon Lavie, 2008). If more than one reference translation is available, the translation is scored against each reference independently, and the best scoring is used.

**Bilingual Evaluation Understudy (BLEU)**

The BLEU method of Mt output scoring was developed in the IBM labs (Papineni et al., 2001) to obtain a rapid and economical way to automatically evaluate machine translation. The initial purpose of designing of this method was to correlate with human assessment. In this scoring method the assessment of MT output is conducted in sentence level.

The basic idea of BLEU is to reward closeness to one of human translation as reference translation, using modified unigram precision (Philipp Koehn, 2010). The precision is determined by the weighted overlap of n-grams from the candidate translation to the reference translation (for n=1, … , 4). The final score between 0 and 1 tells how close the candidate is to reference translation. BLEU is currently the most commonly used score for comparing MT systems and evaluating improvements, because it's easy to compute and provides reasonable performance. In modified n-gram precision, the numerator is bound to the maximum number of occurrences of that n-gram in any other the references.

**Word Error Rate (WER)**

The word error rate is an edit distance measure that originates from the **Levenshtein distance** which is defined as the minimum number of editing steps such as insertions, deletions, and substitutions and uses words instead of characters as basic units (Philipp Koehn, 2010). The WER measures the similarity of two sequences of words by

evaluating the minimum number of deletions, insertion or substitutions needed to turn the candidate sentence into reference (Marcuse Dobrinkat, 2008).

The reliability of evaluation metrics is a highly disputed topic. Although the evaluation result of a metric is correlated with human judgment in most cases, there are still cases it is not. The problem is that there is not a verified understanding about the situations under which a metric will be reliable. For instance, the experiments by Culy and Riehemann (2003) show that two evaluation metrics, i.e., BLEU and NIST, perform poorly to rank the quality of MT output and human translation of literary texts, and some MT outputs can even outscore professional human translations. Callison-Burch et al. (2006) raise an example in a 2005 NIST MT evaluation exercise to show that the system ranked at the top in the human evaluation section is ranked only the sixth by BLEU.

Thurmair (2005) discusses this problem of inconsistent performance from a perspective of the way an evaluation metric judges the translation quality. While the translation quality of an MT output depends on its word similarity towards the corresponding human translation, a direct, word-to-word human translation probably is able to yield a high evaluation score, and a free human translation would then be a disaster. As a whole, Babych et al. (2005) comment that the evaluation metrics cannot give a "universal" prediction of human perception towards translation quality, and their predictive power is "local" to a particular language or text type. For new language pair and text genre, human evaluation of the performance of a metric is necessary to prove its reliability.

Although the performance of evaluation metrics still remains questionable, their use is necessary. When used in the proper way, they are able to show their values. According to Thurmair (2005), automatic metrics are good for ranking different systems and for measuring the overall progress of the same system in different developmental stages, but they are not good for certain scenarios that require in-depth evaluations, for example, the R&D (Research and Development) on system improvement. "To find out that 30% of your errors are unknown words you don't need a BLEU score – and BLEU would not tell you" (Thurmair 2005).

Therefore, automatic evaluation metrics are not supposed to be used to replace human judgment completely. But they have exhibited a great value in making large-scale MT evaluations possible at a controllable cost. It has been a consensus in the field that an automatic MT evaluation with large-scale data sets gives more objective and less biased results than manual evaluation totally relying on subjective and inconsistent judgment on a small set of sentences. In the next chapter the methodology of this study is presented in details.

## Methodology of Study

This study is an attempt to make a comparison between two kinds of common evaluation methods in MT studies namely: Human evaluation and Automatic evaluation of MT systems. At the same time, it is an attempt to investigate performance of two common English to Persian MT systems in practical environment, Google translator system and Pars translator system. Based on Koehn (2010) in manual evaluation the outputs are assess based on correctness standard, whoever correctness is a broad measure of assessment for this it's divided to two sub-criteria as: fluency and adequacy, which based on Philipp Koehn (2010) are defined as follow:

**Fluency**: is the output good fluent Persian (target language of this study)? This involves both grammatical correctness and idiomatic word choices.

**Adequacy**: does the output convey the same meaning as the input sentence? Is part of message lost, added, or distort?

In order to assess collected data automatically; BLUE method is applied based on Philipp Koehn (2010) this method was initiated to correlate with human assessment. The basic idea of BLEU is to reward closeness to one of human translation as reference translation, using modified unigram precision (Philipp Koehn, 2010). The precision is determined by the weighted overlap of n-grams from the candidate translation to the reference translation (for n=1…, 4). The final score between 0 and 1tells how close the candidates are to reference translation.

### Participations

Twenty MA students of translation studies participated in this study. All of these subjects that here referred to them as evaluators, were in their final semester of translation studies course and more or less have the same amount of knowledge in translation and were chosen based on their ability of translating from English to Persian.

### Text Selection

20 sentences were selected from the book of " Translating of Economical Texts" published by Payam-e-Nuor University with taking into consideration all aspects of translation problems including translation of long and short sentences, translation of abbreviations, translation of special expression in economic texts and …etc. Ongoing

need to translating Economical texts in the world was one of the important criteria in selecting this kind of text and in order to avoid unreliable texts it is attempted all samples were selected from one book that is published by credible and academic publication. It is noteworthy to add that the reference translations presented in this study are also from above mentioned book to reduce the problem that might cause by interference of evaluators' various taste and style.

## Instruments

The sentences were translated from English to Persian by two common MT systems (Google translate MT and Pars translation MT). The reasons for selecting Google translate MT system are as follow: Google translation system enjoy the technology of statistical translation method and is easily accessible online 24/7 and is free-of charge, also it is able to translate a lengthy text with different subjects. Google translate is an automatic MT system that works without intervention of human translators.

Another system that is used in this study is Pars translation MT system that translates English text into Persian sentences. Pars system is the first commercial version of software which issued to public in June 1997 and the last up-dated version was delivered in April 2004. As it is claimed by its designers Pars Trans engine is able to recognize and parse for more than 1.5 million words and terminologies commonly used in public English and 33 fields of sciences also its bank of words and terminologies are being reviewed continuously and upgraded by their stuffs in academic centers. In the next chapter the results of the study are presented.

## Result of Study

As it's noted in the previous section, data are collected from economical texts. Translated texts by both MT systems (Google translate and Pars translation system) were assessed by human evaluation and automatic evaluation methods, in the first part of this section the automatic evaluation method (BLEU) is applied and the results are shown by tables, then human evaluation approach is used to evaluate output of both under investigation systems.

### Automatic MT Evaluations

In order to evaluate collected data automatically, the BLEU method of evaluation is selected in this study. This method is presented by Papineni et al. (2001) and is a

language independent metric to provide a quick overview of the performance of an MT system. For measuring the quality of MT this method, follows very simple hypothesis the closer a machine translation is to a professional human translation, the better it is. In practice, BLEU works, by comparing its translation against available reference translation by human translators.

The results of evaluation are presented by following tables:

1. In the walrasian scheme, factors of production are concrete items in existence at a moment of time.

Reference translation:

در طرح والراس: عوامل تولید اقلام مشخصی را تشکیل می دهند که در لحظه ای از زمان وجود دارند.

Candidate 1(Google translation system) 10/ 15= 0.66

، عوامل تولید اقلام بتن در وجود در یک لحظه از زمان است.walrasianدر این طرح

Candidate 2(Pars translation system) 6/ 14= 0.42

walrasian ،عوامل تولید اقلام ملموس در وجود در یک گشتاور زمان هست .در طرح

2. Walras seems deliberately to slur over the distinction between income from work and income from property.

Reference translation:

بنظر می رسد والراس آگاهانه تمایز بین درآمد از کار و درآمد از مالکیت را لوث می کند.

Candidate 1 10/20= 0.2

به نظر می رسد عمدا به بیش از تمایز بین درآمد□ا□ل از کار و درآمد□ا□ل از اموال لکه دار Walras کردن.

Candidate 2 8/21= 0.38

عمدا به نظر می رساند که بر روی تمایز مابین درآمد از کار و درآمد از دارایی در بهم بر هم Walras بنویسد.

3. All factors are free and equal in the market.

Reference translation:

در بازار همه عوامل آزاد و برابرند.

Candidate 1 7/7= 1

همه عوامل در بازار آزاد و برابر هستند.

Candidate 2 6/8= 0.75

تمام عوامل مجانی و برابر در بازار هستند.

4. Economic activity was organized on the assumption of cheap and abundant oil.

Reference translation:

فعالیتهای اقتصادی را بر اساس نفت ارزان و فراوان سازماندهی می کردند.

Candidate 1 6/12= 0.5

فعالیتهای اقتصادی در این فرض از نفت ارزان و فراوان برگزار شد.

Candidate 2 6/12= 0.5

فعالیت اقتصاد روی فرض از ارزان و نفت فراوان سازمان داده شد.

5. We say that the economy is experiencing inflation.

Reference translation:

میگوییم که اقتصاد دچار تورم شده است.

Candidate 1 4/9= 0.44

ما می گویند که اقتصاد در □ال تجربه تورم است.

Candidate 2 5/7= 0.71

ما می گوییم اقتصاد که تورم تجربه است.

6. How does the oil price increase affect what is being produced?

Reference translation:

افزایش بهای نفت چگونه بر نوع تولیدات اثر می گذارد؟

Candidate 1 6 /10= 0.6

چگونه افزایش قیمت نفت تاثیر می گذارد در □ال تولید است؟

Candidate 2 5/11= 0.45

افزایش قیمت نفت چطور متاثر می کند آنچه که ارائه داده می شوید؟

7. The introduction of new products.

Reference translation:

ورود محصولات جدید به بازار.

Candidate 1 2/4= 0.50

معرفی محصولات جدید است.

Candidate 2 2/3= 0.66

معرفی محصولات جدید.

8. Mismatch of skills and job opportunities.

Reference translation:

عدم تناسب بین مهارتهای کاری و فر□تهای شغلی.

Candidate 1 6/6 =1

عدم تناسب مهارت ها و فر□ت های شغلی.

Candidate 2 3/5 = 0.6

Mismatch مهارتها و فر□تهای کار

9. To restore aggregate demand to its full-employment.

Reference translation:

بازگرداندن تقاضای کل به سطح اشتغال کامل.

Candidate 1 5/9 = 0.55

برای بازگرداندن تقاضای کل به تمام اشتغال خود را.

Candidate 2 2/8= 0.25

به تقاضای انباشته به. full - employment خودش برمی گرداند

10. Economic scarcity requires people to make economic choices.

Reference translation:

کمیابی اقتصادی مردم را ملزم میکند که انتخابهای اقتصادی انجام دهند.

Candidate 1 6/8= 0.75

کمیابی اقتصادی مردم را به اختیار انتخابهای اقتصادی.

Candidate 2 4/9= 0.44

کمیابی اقتصاد مردم لازم دارد که انتخابهای اقتصاد بسازد.

11. What are the necessary requirements for a commodity to have such a wide market?

Reference translation:

چه شرایطی لازم است تا کالایی به چنین بازار گسترده ای برسد؟

Candidate 1 6/10= 0.6

شرایط لازم برای یک کالا به چنین بازار گسترده ای چه هستند؟

Candidate 2 3/13= 0.23

خواسته ها لازم برای یک کالا چه چیزی هستند که چنان یک بازار وسیع دارد؟

12. Unless his profits were to bear some proportion to the extent of his stock.

Reference translation:

مگر اینکه سود او به نسبت سرمایه اش بالا برود.

Candidate 1 4/15= 0.27

مگر در مواردی که سود خود را به تحمل نسبت به میزان سهام خود را.

Candidate 2 4/13= 0.30

مگر اینکه سودهایش بودند که مقداری تناسب به وسعت موجودش پیش فروش بکنند.

13. It improved the climate for investment and so help to maintain aggregate demand.

Reference translation:

عقیده بر این بود که این افزایش شرایط سرمایه گذاری را بهبود می بخشد و بدین ترتیب به□فظ میزان تقاضای کل کمک میکند.

Candidate 1 4/13= 0.30

. بهتر آب و هوا برای سرمایه گذاری و کمک به □فظ تقاضای کل است.

Candidate 2 3/15= 0.2

آن آب و هوا برای سرمایه گذاری و بنابراین کمک به تقاضای انباشته نگهداری ا□لاح کرد .

14. Buoyant sellers' market.

Reference translation:

بازار فروش پر رونق.

Candidate 1 1/5= 0.2

در بازار فروشندگان شناور است.

Candidate 2 1/4= 0.25

بازار داغ بازارفروشنده ها.

15. Excessive demand causes rising prices and balance of payments difficulties.

Reference translation:

افزودگی تقاضا موجب افزایش قیمتها و بحران بیلان پرداختها می شود.

Candidate 1 5/11= 0.45

تقاضای بیش از □د باعث افزایش قیمت ها و تعادل مشکلات پرداخت.

Candidate 2 4/12= 0.33

.تقاضا بیش از □د قیمتهای خیزان و مشکلهای تراز پرداخت ها سبب می شود

16. GDP and GNP measure the total output and total income of an economy.

Reference translation:

تولید کلی و درآمد کلی یک اقتصاد بر اساس تولید ناخالص داخلی و تولید ناخالص ملی اندازه گیری می شود.

Candidate 1 7/14= 0.5

تولید ناخالص داخلی و تولید ناخالص ملی اندازه گیری خروجی کل و درآمد کل اقتصاد است.

Candidate 2 6/14= 0.43

GDP GNP تولید تمام را و درآمد تمام از یک اقتصاد اندازه می گیرد .

17. Exploitation of economies of scale.

Reference translation:

بهره برداری از □رفه جویی های مقیاسی.

Candidate 1 2/4= 0.5

بهره برداری از اقتصاد مقیاس.

Candidate 2 1/3= 0.33

بهره کشی □رفه جوییهای مقیاس.

18. Current market share.

Reference translation:

سهم جاری از فروش بازار.

Candidate 1 2/3= 0.67

سهم بازار کنونی.

Candidate 2 3/3= 1

سهم بازار جاری.

19. Oil-related commodities.

Reference translation:

کالاهای وابسته به نفت.

Candidate 1 2/4= 0.5

مربوط به کالاهای نفت.

Candidate 2 1/3= 0.33

کالاها. Oil – related

20. Positive economics deals with objective or scientific explanations of the working of the economy.

Reference translation:

اقتصاد اثباتی با مصداق های عینی و علمی چگونگی عملکرد اقتصاد سرو کار دارد.

Candidate 1 8/11= 0.72

معاملات مثبت اقتصاد با توضیحات عینی و علمی از کار اقتصاد.

Candidate 2 5/12= 0 .42

اقتصاد اثباتی با هدف یا تبیینهای علمی از کاری اقتصاد معامله می کند.

Based on the results of evaluation candidate no. 1 (Google translate system) has gained higher BLEU score with sum of 10.91 against candidate no. 2 (pars translation system) with sum of 8.98 scores.

**Human Evaluation**

In order to Evaluate performance of the both Google translate and Pars systems manually adequacy and fluency of translated text were analyzed based on human evaluation method presented in (Philipp Koehn, 2010) in this method both criteria of evaluation are scored from 1 to 5 based on their quality which in this study is assessed by 10 MA translation students. The details of the evaluation are depicted in the following table:

Table 4.1 obtained results of MT outputs by BLEU methods of evaluation

| System / sentences | Google | Pars | System / sentences | Google | Pars |
|---|---|---|---|---|---|
| Sentence 1 | 0.66 | 0.42 | Sentence 11 | 0.6 | 0.23 |
| Sentence 2 | 0.2 | 0.38 | Sentence 12 | 0.27 | 0.30 |
| Sentence 3 | 1 | 0.75 | Sentence 13 | 0.30 | 0.2 |
| Sentence 4 | 0.5 | 0.5 | Sentence 14 | 0.2 | 0.25 |
| Sentence 5 | 0.44 | 0.71 | Sentence 15 | 0.45 | 0.33 |
| Sentence 6 | 0.6 | 0.45 | Sentence 16 | 0.5 | 0.43 |
| Sentence 7 | 0.50 | 0.66 | Sentence 17 | 0.5 | 0.33 |
| Sentence 8 | 1 | 0.6 | Sentence 18 | 0.67 | 1 |
| Sentence 9 | 0.55 | 0.25 | Sentence 19 | 0.5 | 0.33 |
| Sentence 10 | 0.75 | 0.44 | Sentence 20 | 0.72 | 0.42 |

The output of both systems are scored on a scale of 1-5 for fluency (good English) and adequacy (correct meaning) and are described as fellow: (Philipp Koehn, 2010)

Table 4.2 scoring Adequacy and Fluency in current study based on Philipp Koehn model (2010)

| score | Adequacy | score | Fluency |
|---|---|---|---|
| 5 | all meaning | 5 | flawless Persian |
| 4 | most meaning | 4 | good Persian |
| 3 | much meaning | 3 | non- native Persian |
| 2 | little meaning | 2 | disfluent Persian |
| 1 | none | 1 | incomprehensible |

As it is noted in previous chapter translated texts by MT systems are analyzed by ten evaluators. In the following tables the average of scores is presented for each system separately.

Table 4.3 assessing Adequacy and Fluency of Google translate system

| Fluency | | | | | adequacy | | | | | Translation by Google | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 4 | 3 | 2 | 1 | 5 | 4 | 3 | 2 | 1 | | |
| | | | * | | | | * | | | ، عوامل تولید اقلام بتن در walrasianدر این طرح وجود در یک لحظه از زمان است. | 1 |
| | | * | | | | * | | | | به نظر می رسد عمدا به بیش از تمایز بین درآمد Walras □ا□ل از کار و درآمد□ا□ل از اموال لکه دار کردن. | 2 |
| | | * | | | | * | | | | همه عوامل در بازار آزاد و برابر هستند. | 3 |
| | | | * | | | | * | | | فعالیتهای اقتصادی در این فرض از نفت ارزان و فراوان برگزار شد. | 4 |
| | | * | | | | | * | | | ما می گویند که اقتصاد در □ال تجربه تورم است. | 5 |
| | | | * | | | | * | | | چگونه افزایش قیمت نفت تاثیر می گذارد در □ال تولید است؟ | 6 |
| | | * | | | | * | | | | معرفی محصولات جدید است. | 7 |
| | | * | | | | * | | | | عدم تناسب مهارت ها و فر□ت های شغلی. | 8 |
| | | | * | | | | | * | | برای بازگرداندن تقاضای کل به تمام اشتغال خود را. | 9 |
| | | | * | | | | * | | | کمیابی اقتصادی مردم را به اختیار انتخابهای اقتصادی. | 10 |
| | | | * | | | | * | | | شرایط لازم برای یک کالا به چنین بازار گسترده ای چه هستند؟ | 11 |
| | | | * | | | | | * | | مگر در مواردی که سود خود را به تحمل نسبت به میزان سهام خود را. | 12 |
| | | * | | | | | * | | | بهتر آب و هوا برای سرمایه گذاری و کمک به □فظ تقاضای کل است. | 13 |
| | | * | | | | | * | | | در بازار فروشندگان شناور است. | 14 |
| | | * | | | | * | | | | تقاضای بیش از □د باعث افزایش قیمت ها و تعادل مشکلات پرداخت. | 15 |
| | | | * | | | | * | | | تولید ناخالص داخلی و تولید ناخالص ملی اندازه گیری خروجی کل و درآمد کل اقتصاد است. | 16 |
| | | * | | | | | * | | | بهره برداری از اقتصاد مقیاس. | 17 |
| | | | * | | | | * | | | سهم بازار کنونی. | 18 |
| | | * | | | | * | | | | مربوط به کالاهای نفت. | 19 |
| | | * | | | | | * | | | معاملات مثبت اقتصاد با توضیحات عینی و علمی از کار اقتصاد. | 20 |

In order to analyze collected data through Human MT evaluation method, SPSS software is used, the following tables illustrate mean and standard deviation of adequacy

and fluency between Google and Pars MT system. In table no. 4.5 the maximum and minimum of acquired scores are also presented and the scores of MT systems are compared with human translation (reference translation). Because reference translation, in this study, is considered as the reference of comparison, full score (5) is dedicated to it as the best and standard translation of the sentences.

Table 4.4 assessing Adequacy and Fluency of Pars translate system

| Fluency | | | | | adequacy | | | | | Translation by Pars system | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 4 | 3 | 2 | 1 | 5 | 4 | 3 | 2 | 1 | | |
| | | | | * | | | * | | | ،عوامل تولید اقلام ملموس در وجود در walrasianدر طرح یک گشتاور زمان هست | 1 |
| | | | | * | | | | | * | عمدا به نظر می رساند که بر روی تمایز مابین درآمد از Walras کار و درآمد از دارایی در بهم بر هم بنویسد. | 2 |
| | | * | | | | | * | | | تمام عوامل مجانی و برابر در بازار هستند | 3 |
| | | | * | | | | | * | | فعالیت اقتصاد روی فرض از ارزان و نفت فراوان سازمان داده شد. | 4 |
| | | | * | | | | | | * | ما می گوییم اقتصاد که تورم تجربه است | 5 |
| | | | * | | | | | * | | افزایش قیمت نفت چطور متاثر می کند آنچه که ارائه داده می شوید؟ | 6 |
| | | * | | | | | * | | | معرفی محصولات جدید. | 7 |
| | | * | | | | | | * | | Mismatch مهارتها و فر□تهای کار | 8 |
| | | | | * | | | | | * | به تقاضای full - employment خودش برمی گرداند انباشته به. | 9 |
| | | | * | | | | | * | | کمیابی اقتصاد مردم لازم دارد که انتخابهای اقتصاد بسازد. | 10 |
| | | * | | | | | * | | | خواسته ها لازم برای یک کالا چه چیزی هستند که چنان یک بازار وسیع دارد؟ | 11 |
| | | | | * | | | | | * | مگر اینکه سودهایش بودند که مقداری تناسب به وسعت موجودش پیش فروش بکنند. | 12 |
| | | | | * | | | | | * | آن آب و هوا برای سرمایه گذاری و بنابراین کمک به تقاضای انباشته نگهداری ا□لاح کرد . | 13 |
| | | * | | | | | * | | | بازار داغ بازارفروشنده ها. | 14 |
| | | | * | | | | | * | | تقاضا بیش از□د قیمتهای خیزان و مشکلهای تراز پرداخت ها سبب می شود. | 15 |
| | | | * | | | | | * | | تولید تمام را و درآمد تمام از یک اقتصاد اندازه GDP و GNP می گیرد. | 16 |
| | | * | | | | | * | | | بهره کشی □رفه جوییهای مقیاس. | 17 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | * | | | | | * | | سهم بازار جاری | 18 |
| | | | * | | | | | * | | کالاها. Oil – related | 19 |
| | | | | * | | | | | * | اقتصاد اثباتی با هدف یا تبیینهای علمی از کاری اقتصاد معامله می کند. | 20 |

In order to analyze collected data through Human MT evaluation method, SPSS software is used, the following tables illustrate mean and standard deviation of adequacy and fluency between Google and Pars MT system. In table no. 4.5 the maximum and minimum of acquired scores are also presented and the scores of MT systems are compared with human translation (reference translation). Because reference translation, in this study, is considered as the reference of comparison, full score (5) is dedicated to it as the best and standard translation of the sentences.

In table no. 4.6 performance of MT systems are compared in order to determine more reliable MT system and also making judgment about which MT system's output is closer to reference translation.

Table 4.5 mean and standard deviation of adequacy of MT systems' output.

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Reference | 20 | 5.00 | 5.00 | 5.0000 | .00000 |
| Adequacy of Google | 20 | 2.00 | 4.00 | 3.1500 | .58714 |
| Adequacy of Pars | 20 | 1.00 | 3.00 | 2.1500 | .81273 |

Table 4.7 mean and standard deviation of fluency of MT systems' output

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Fluency of reference sentences | 20 | 5.00 | 5.00 | 5.0000 | .00000 |
| Fluency of Google output | 20 | 2.00 | 3.00 | 2.4000 | .50262 |
| Fluency Pars output | 20 | 1.00 | 3.00 | 2.0000 | .79472 |

In the next chapter the results reported in chapter 4 will be discussed at length.

## Discussion

In order to evaluate outputs of MT system automatically BLEU method is applied. Based on this method, the closer a machine translation is to a professional human translation, the better it is. Generally speaking, BLEU works by comparing MT output against available reference translation by human translators. According to the results of evaluation illustrated by table 4.1 candidate no. 1 (Google translate system) has gained higher BLEU score with sum of 10.91 against candidate no. 2 (pars translation system) with sum of 8.98 scores.

To evaluate outputs humanly variables of adequacy and fluency were assessed based on the model of Philipp Koehn (2010). The outputs of both systems are scored on a scale of 1 to 5 for adequacy (correct meaning) and fluency (good Persian) as it's depicted by table 4.2.

At a quick glance to the tables 4.5 and 4.7, it can be elicit that performance of Google translate system in variable of adequacy with mean of 3.15 and Std. deviation of 0.58 is better than adequacy of Pars system with mean of 2.15 and Std. deviation of 0.81. In analyzing variable of fluency between two systems the same results repeated where the fluency of Google translate system's outputs are better than Pars system's outputs with mean and Std. deviation of 2.4±0.50 and 2±0.79; therefore, it can be concluded that performance of Google system with higher mean and lower Std. deviation is better than performance of Pars system. In order to analyze these data in details, they are analyzed by SPSS software it should be noted that the mean difference is considered significance at the level of .05and as it is illustrated in table 4.9 the following results were obtained:

1. With regard to variable of adequacy, there is a significant difference between reference translation and Google output, since P-value = .000<0.05

2. There is a significance difference between reference translation and output of Pars MT system because P- value = .000<0.05.

3. There is a significance difference between adequacy of Google translate output and output of Pars system, since P- value = .000<0.05.

As it is illustrated by table 4.10, obtained results pertinent to the variable of fluency show that:

1. There is a significance difference between reference translation and output of Google translate MT system, since P- value = .000<0.05.

2. There is a significance difference between reference translation and output of Pars MT system, since P- value = .000<0.05.

3. But, there is no significance difference between fluency of output of Google translate MT system and Pars system's output because P- value = 0.65>0.05

## Conclusion

As it is discussed in previous section, the main concern of this study is to make a comparison between human evaluation and automatic evaluation of MT systems, at the same time; performances of two English to Persian MT system were analyzed through both methods of evaluations.

In order to evaluate collected data automatically the model of BLEU is used. According to this method translated text by MT systems should be compared with one or more human translations (reference translations) of the same text. Based on the findings of the study Google translate system has a better performance than Pars system (research question No. 3) , this result obtained by analyzing data through bout method of MT evaluation (human and automatic evaluation method), then in order to answer the research questions No. 1 and 2 these two methods of MT evaluation lead to the same results but human evaluation which is done by professional translators is more reliable than automatic translation; however, it is more expensive and time consuming than automatic MT evaluation methods.

## References

Aal Attia, M.A. (2002) Implications of the Agreement Features in Machine Translation. Master Thesis, University of Al- Azhar , Cairo , Egypt.

Agarwal, A & Lavie , A . (2008). METEOR, MBLEU and M-TER: Evaluation Metric for High Correlation with Human Rankings of Machine Translation Output. In Proceedings of the Third ACL Workshop on Statistical Machine Translation, Columbus, USA, Jun

Allen B. Tucher, (1987) "Current strategies in machine translation research and development", in Sergei Nirenburg, ed., Machine Translation: Theoretical and methodological issues, Cambridge: Cambridge University Press.

ALPAC. (1966). Language and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee. Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences - National Research Council.

Arnold, D, at el. (1994) Machine Translation: An Introductory Guide. Manchester: Blackwell.

Arnold, D, Sadler, L. and Humphreys, R. L. (1993). Evaluation: An Assessment. Machine Translation. Vol. 8, Number 1-2. Special issue in Evaluation of MT

Arturo, T. (1999) Translation Engines: Techniques for Machine Translation. London: Springer

Babych, B at el. (2005). Estimating the predictive power of n-gram MT evaluation metrics across language and text types. In *MT Summit X.* Phuket, 13-15 September 2005.

Beeby, A.Rodriguez Ines, P. & Sanchez–Gijon, P. (Eds) (2009). Corpus Use and Translating, John Benjamin Publication Company

Callison-Burch, C at el. (2006). Re-evaluating the Role of Bleu in Machine Translation Research.

Carrol, J. B . (1966). An experiment in evaluating the quality of translation.National Academy of Sciences National Research Council Washington, D. C.

Chakaveh Saedi at el. (2009). Automatic Translation between English and Persian texts. CAASL-3 – Third Workshop on Computational Approaches to Arabic Script-based Languages [at] MT Summit XII, August 26, 2009, Ottawa, Ontario, Canada.

Crook & Bishop (1979). Measurement of readability by the cloze test. In van Slype, G (1979). Critical Methods for Evaluating the Quality of Machine Translation. Prepared for European Commission. Report BR 19142.

Culy, C. and Riehemann, S. Z. (2003). The limits of n-gram translation evaluation metrics *MT Summit IX*, New Orleans, USA, 23-27 September 2003.

Systems

Trujillo, A (1999). Translation Engines: Techniques for Machine Translation, London: Springer.