Advances in Language and Literary Studies
ISSN: 2203-4714
Vol. 6 No. 4; August 2015
Australian International Academic Centre, Australia
Flourishing Creativity & Literacy

An Evaluation of Output Quality of Machine Translation (Padideh Software vs. Google Translate)

Haniyeh Sadeghi Azer
Department of Translation Studies, East Azerbayjan Science and Research Branch, Islamic Azad University, Tabriz, Iran
E-mail: Haniyeh.sadeghiazar@gmail.com

Mohammad Bagher Aghayi (Corresponding author)
Department of Translation Studies, East Azerbayjan Science and Research Branch, Islamic Azad University, Tabriz, Iran
E-mail: aghaeimb@yahoo.com

Doi:10.7575/aiac.alls.v.6n.4p.226
URL: http://dx.doi.org/10.7575/aiac.alls.v.6n.4p.226
Received: 10/04/2015
Accepted: 30/06/2015

Abstract

This study aims to evaluate the translation quality of two machine translation systems in translating six different text-types, from English to Persian. The evaluation was based on criteria proposed by Van Slype (1979). The proposed model for evaluation is a black-box type, comparative and adequacy-oriented evaluation. To conduct the evaluation, a questionnaire was assigned to end-users to evaluate the outputs to examine and determine, if the machine-generated translations are intelligible and acceptable from their point of view and which one of the machine-generated translations produced by Padideh software and Google Translate is more acceptable and useful from the end-users point of view. The findings indicate that, the machine-generated translations are intelligible and acceptable in translating certain text-types, for end-users and Google Translate is more acceptable from end-users point of view.

Keywords: Machine Translation, Machine Translation Evaluation, Translation Quality

1. Introduction

Language is a tool for communication. Every people, all over the world need a language to communicate with others. Sometimes people do not know each other's language, so it gets impossible to interact with each other. In those cases, a person or a tool is needed to translate the source Language into the target language. From earlier times till now, human translators, translate speech and documents in a foreign language and help people to understand each other.

However, human translators are not always available and easy to find. Also, the amount of written material that a person can translate in a specific time is very limited. The translation process is very time consuming, and moreover, having a human translator is costly. Therefore, searching for alternative methods for translation is crucial.

By the emergence of computers, the idea of using them in the automatic translation process developed. Using computers for translation proposes a solution for all these costly and time consuming processes which have to be done by human translator. Machine translation's purpose is to reduce the cost of the translation process and increase the quality of the translated material.

However, translating a language into another one through computer is not an easy task. A human language is a very complicated system, so Machine translation involves a great deal of complicated analysis and manipulation, and despite the advances that are done in this field but it is not accomplished yet.

The evaluation of machine translation systems is an important field of research, for optimizing the performance of MT systems and their effectiveness. There are a range of different evaluation approaches for evaluating MT systems; progress in the field of machine translation relies on assessing the quality of a new system through systematic evaluation. The evaluation strategy adopted in this study is human evaluation, and the focus is on manual corpus analysis and human judgments on machine-generated translation. It intends to report an evaluation of the output quality of two prevalent English-Persian MT programs, named, Padideh software and Google translate. The aim of the research is to find out which program produces a relatively better output, in dealing with diverse text-types in translation direction from English to Persian, and its acceptability and usability for end-users.

The use of MT or any sort of computerized tool for translation support is completely unknown to the vast majority of individuals and organizations, even those involved in the 'language industries', like translators.

Most of the time, users of MT cannot select proper MT systems compatible to their needs and their purpose for using MT. Arnold, et al (1994) indicates that the purchase of an MT system is in many cases a costly affair and requires careful consideration. It is important to understand the organizational consequences and to be aware of the system's capacities. Evaluation of MT systems helps to inform about the usability and acceptability of them.

The present study raises questions, regarding the evaluation of the two so-called MT systems, and aims to investigate the following main questions:

RQ1: Are machine-generated translations intelligible and acceptable from the point of view of end-users of diverse text-type of documents?

RQ2: Which one of the machine-generated translations produced by Padideh software and

Google translate is more acceptable and useful from the end-users point of view?

The Aims of present study is to establish whether six different text types target language translations produced by two prevalent machine translation softwares (Google translate and Padideh translator) are considered intelligible and acceptable from the point of view of end- users (RQ1), and which one of the machine-generated translations produced by them is more acceptable and useful from their point of view (RQ2). These research questions are investigated through human evaluation of machine translation output. Therefore in order to meet the aims proposed, the study developed to use a human evaluation model to conduct end-user evaluations of diverse text-types.

2. Methodology

2.1 Overview

The research design, employed in this study is build on previous work conducted by Van Slype (1979). Criteria of evaluation are established by Georges Van Slype (1979) Method for evaluating the quality of Machine Translation from the perspective of acceptance and usability for the end-users.

Two English-Persian Machine translation program (Padideh software & Google Translate) are selected as the subject of this research. The research only evaluates the output quality of Machine translation programs. Different text-types have been selected, in order to examine the translation produced by each program.

2.2 Theoretical Framework

The evaluation made in this research focused on the quality of the output, i.e., the translation of two prevalent English-Persian MT programs. The evaluation of these two different translation programs will be established by implementing Van Slype (1979) method, for evaluating machine translation.

In 1979, Van Slype compiled a comprehensive critical review of MT evaluation methods on behalf of Bureau Marcel van Dijk for the Commission of the European Communities, who had set up a program aimed at "lowering the barriers between the languages of the Community" (Van Slype, 1979, p.11). The purposes of this study were: to document the kinds of methodologies being employed at this time in MT evaluation; to make some recommendations to the Commission, amongst other things, on the methodology it should use when evaluating its machine translation systems; and to conduct research which would help in the long term with the efficiency of these evaluations. The report distinguished between two levels of evaluation: macroevaluation (or total evaluation) determines the acceptability of a system, compares the quality of two systems or two versions of the same system, and assesses the usability of a system; while microevaluation (or detailed evaluation) determines the improvability of a system.

2.2.1 Macroevaluation

This level of evaluation concerns itself with the assessment of the system's overall performance (Van Slype, 1979, p. 88). It deals with all the criteria and all the methods used or proposed to assess the "static" quality of an MT system, i.e., its quality at the moment of evaluation, and regardless of the manner by which this quality has been reached. It aims at examining the acceptance of a translation system, comparing the quality of two translation systems or two versions of the same system and/or assessing the usability of a translation system (Van Slype, 1979, pp.12 and 21).

Van Slype (1979, p.56) points out that the macroevaluation of a system is the operation which consists in assessing the manner in which the system answers the requirements and the needs of its users, actual or potential, regardless of what occurs inside the "black-box". It has the purpose of measuring the adequacy of the output from the system to its environment, without seeking to diagnose the causes of its inadequacy, if any, and without pinpointing the component(s) that could usefully be modified to improve adequacy.

Van Slype (1979) broke down the various criteria into ten classes, assembled in turn into four groups according to the level at which they approach the quality of the translation.

- Cognitive level (effective communication of information and knowledge).
    - Intelligibility
    - Fidelity
    - Coherence
    - Usefulness
    - Acceptability
- Economic level (excluding costs).
    - Reading time
    - Correction time
    - Translation time

- Linguistic level (conformity with a linguistic model)

- Operational level (effective operation).

Description of criteria and methods of macroevaluation, used in this study:

Cognitive level:

1. Intelligibility:

Van Slype (ibid) defines the criteria as:

Subjective evaluation of the degree of comprehensibility and clarity of the translation.

Van Slype (ibid): Measurement of intelligibility by rating sentences on a 4-point scale.

* Method:

-Submission of a text sample in several versions (original text, MT without and with post-editing, human translation with and without revision) to a group of evaluators; the texts being distributed so that each evaluator:

· Receives only one of each of the versions of the texts.

· Receives a series of sentences in sequence (sentences in their context).

-Rating of each sentence according to a 4-point scale.

-Calculation of the average of the ratings per text and version, with and without weighting as a function of the number of words in each sentence evaluated.

* Scale of intelligibility:

3: Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language.

2: Fairly intelligible: the major part of the message passes.

1: Basely intelligible: a part only of the content is understandable, representing less than 50% of the message.

0: Unintelligible: nothing or almost nothing of the message is comprehensible.

2. Fidelity:

Van Slype (ibid), defines fidelity as:

Subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation.

The fidelity rating should, generally, be equal to or lower than the intelligibility rating, since the unintelligible part of the message is of course not found in the translation. Any variation between the intelligibility rating and the fidelity rating is due to additional distortion of the information, which can arise from:

-A loss of information (silence) (example: word not translated).

-Interference (noise) (example: word added by the system).

-A distortion from a combination of loss and interference (example: word badly translated).

Measurement of fidelity by rating on a 4-point scale:

* Method:

-Submission of a sample of original texts, with the corresponding translations, to one or more evaluators.

-Successive examination of each sentence, in the first place in the translation, then in the original text.

-Rating of the fidelity, sentence by sentence.

-Calculation of the average of the fidelity ratings.

* Scale of fidelity:

3: Completely or almost completely faithful.

2: Fairly faithful: more than 50 % of the original information passes in the translation.

1: Barely faithful: less than 50 % of the original information passes in the translation.

0: Completely or almost completely unfaithful.

3. Coherence:

One author only, Y. WILKS (cited in Van Slype 1979), proposes this criterion:

* Definition of the criterion:

The quality of a translation can be assessed by its level of coherence without the need to study its correctness as compared to the original text. Once a sufficiently large sample is available, the probability that the translation should be at the same time coherent and totally wrong is very weak.

* Method of evaluation:

Y. WILKS does not indicate, unfortunately, how in practice it is possible to rate the coherence of a text. He notes that if an original text may be coherent; this means that any assessment of the coherence of its MT version may not be absolute, based on the MT, but must be relative, as compared to the coherence of the source text. But then one is once again compelled to use bilingual evaluators.

4. Usability:

Definition of the criterion:

One author, W. LENDERS (cited in Van Slype 1979), defines usability (which he also calls applicability) as the possibility to make use of the translation.

Another, P. ARTHERN (cited in Van Slype 1979), defines usability as far as a translation service is concerned, as revisibility.

* Method:

B.H.Dostert (ibid): Measurement of the quality by direct questioning of the final users.

5. Acceptability:

Definition of the criterion:

Van Slype defines acceptability as "a subjective assessment of the extent to which a translation is acceptable to its final user" (ibid, p.92). Van Slype maintains that acceptability can be effectively measured only by a survey of final users and this is illustrated in his suggested subjective evaluation, the second of two methods for evaluating acceptability in the report:

1. Measurement of acceptability by analysis of user motivation, and
2. Measurement of acceptability by direct questioning of users.

Measurement of acceptability by direct questioning of users:

* Method:

- Submission of a sample of MT with the original texts and the corresponding HTs, to a sample of potential users.

- Questions asked (among others).

· Do you consider the translation of these documents to be acceptable, knowing that it comes from a computer and that it can be obtained within a very short time, of the order of half a day?

* In all cases.

* In certain circumstances (to be specified).

* Never.

* For myself.

* For certain of my colleagues.

· Would you be interested in having access to a system of machine translation providing texts of the quality of those shown to you?

6. Reading time:

Reading time can be assessed in various ways:

Van Slype (ibid): by timing the time spent by the evaluator in reading each text of the sample.

2.3 The Corpus

The corpus selected for this study is, six different text types which are selected for English to Persian MT and evaluation. The different text-types are: 1) Kid's Story 2) Political Text 3) Computer Science Text 4) Legal Text 5) A Poem as a Literary Text 6) A Webpage.

The corpus selected for the study is six complete texts, that haven't separated from their context.

The SL texts have been collected from university textbooks and Internet websites. Most of these texts have been selected on the basis of being rich in domain-specific terminologies.

Each of the sample texts has translated once by Padideh software and once by Google Translate.

2.4. Research Methodology and Approaches

In this study we employ a quantitative research design, that using this approach enables a better understanding of research problems.

For a study to be valid it must be reliable. For this study it is essential that valid and reliable measurement techniques are developed and employed when collecting and analysing the data. Relating this to our study of MT evaluation, we must ensure the corpus analysis techniques in the interview questionnaire design are valid and reliable. In order to

minimize errors, we systematically conduct the analysis on the corpus, and for the design of the interview questionnaire we build on the work of Van Slype (1979).

The proposed model for the functional attributes is a black-box type superficial, comparative and adequacy-oriented evaluation. In other words, there is no interaction with the systems tested and the goal is to determine whether output is actually helpful to the user groups in question.

On the basis of the tasks relevant to the end-user's needs in this study, only six functional quality characteristics have been investigated. These include: 'intelligibility', 'fidelity', 'coherence', 'usability', 'acceptability' and 'reading time'. In this work, the black-box evaluation has been chosen due to the fact that commercial MT systems can only be evaluated by this approach (Volk, 2001). Consequently, there has been no access to the inner workings of these systems. Even so, it is desirable to be able to draw from such an evaluation enough conclusions about the various system components.

3. Data Analysis

3.1 Overview

This section discusses the data analysis and findings of the study. Detailed analyses and classifications of the results concerning the various criteria types are presented with tables and charts. The questionnaire used in this study was carefully analysed to ensure that the data gathered was presented clearly.

A detailed analysis based on the black-box approach, superficial and adequacy/ declarative evaluation of six various text types for each of the two MT systems reveals the results.

These results are classified and presented on the basis of:

- Variation in scores between raters.
- Comparison of systems for text types.
- Average of scores of raters.
- Percentage of scores of raters.

The result of the application of the evaluation methods in testing the criteria, take into consideration the grades on the scoring scale, the total score value, and the average score value with respect to each rater and each of the tested MT systems. The evaluation results are reported in tables, which show the distribution of the scores obtained from the investigation of text-types for each of the quality characteristics and MT systems.

3.2 Analysis and Classification of Results

This part is the most important process, which is to calculate the human judgments based on the assigned questionnaire. The evaluators were asked to consider each text and its machine translated outputs to examine the parameters which are provided in the questionnaire. The scores assigned to each parameter by evaluators are shown in Tables and for better analysis; the results are presented in charts for each parameter.

3.2.1 Intelligibility

Analysing the scores rated by sixteen evaluators in testing the intelligibility of six different text types for translations produced by each system, results in information showed in Table 3.1 and Figure 3.1. The average quality for intelligibility of all six text-type for Padideh is 23.33% and for Google Translate is 47.77%.

Table 3.1 Intelligibility

		Padideh	Google translate
Intelligibility	Kid's Story	1	1.5
	Political Text	0.1	1.6
	Computer Science Text	1.2	1.9
	Legal Text	0.1	0.9
	Poem	0.6	0.9
	Webpage	1.2	1.8

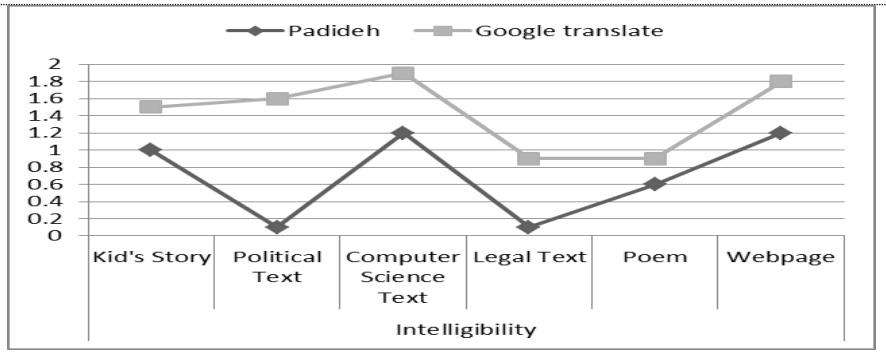

Figure 3.1

3.2.2 Fidelity

After analysing the data reported by evaluators, in evaluating the fidelity of six different text types for translations produced by each system, results in information showed in Table 3.2 and Figure 3.2. The average quality for fidelity of all six text-type for Padideh is 29.22% and for Google Translate is 49%.

Table 3.2 Fidelity

		Padideh	Google translate
Fidelity	Kid's Story	1	1.56
	Political Text	0.69	1.44
	Computer Science Text	1.19	1.75
	Legal Text	0.5	1.13
	Poem	0.75	1.19
	Webpage	1.13	1.75

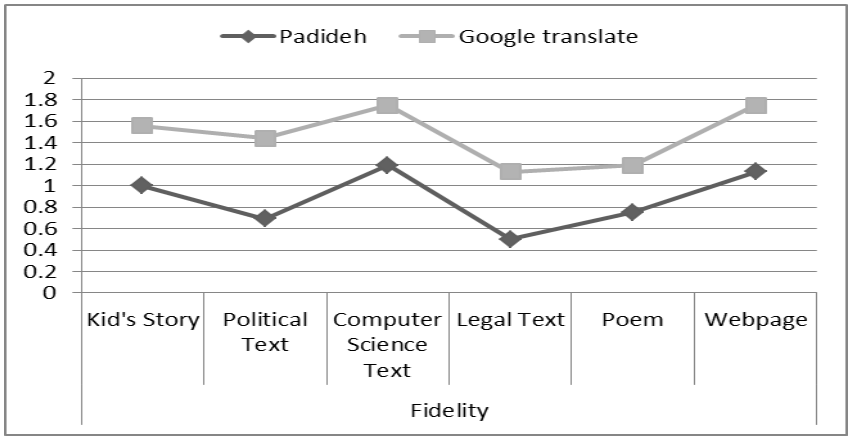

Figure 3.2

3.2.3 Coherence

Exploring the scores rated by sixteen evaluators in examining the coherence of six different text types for translations produced by each system, results in information showed in Table 3.3 and Figure 3.3. The average quality for coherence of all six text-type for Padideh is 36.66% and for Google Translate is 55%.

Table 3.3 Coherence

		Padideh	Google translate
Coherence	Kid's Story	1.1	1.7
	Political Text	1	1.7
	Computer Science Text	1.3	1.9
	Legal Text	0.9	1.1
	Poem	1.4	2.2
	Webpage	0.9	1.3

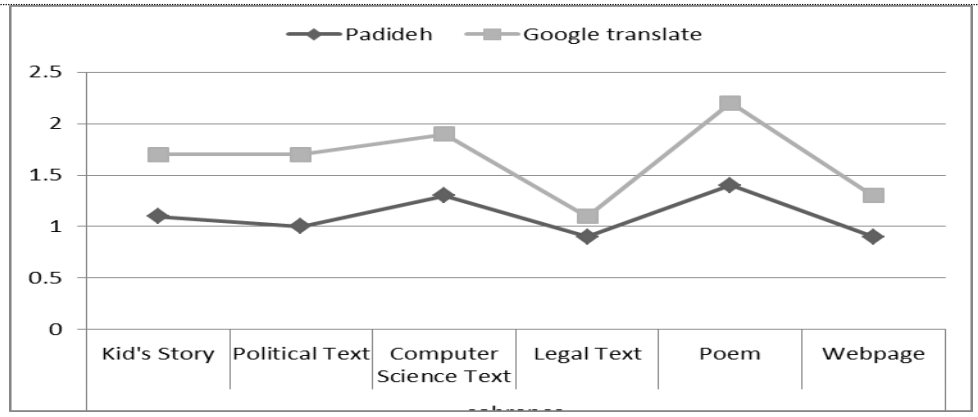

Figure 3.3

3.2.4 Acceptability

For evaluating the acceptability of the systems under investigation, the evaluators were asked, two questions. The evaluators have to answer these questions for each text-type, translated by each system. The analysis of answer of evaluators, have been reported in the base of options number. The collected data from the answers of the evaluators from questionnaire, in evaluating the Acceptability of six different text types for translations produced by Padideh, reveals the information showed in Table 3.4 and Figure 3.4.

Table 3.4 Acceptability(Padideh)

Acceptability(Padideh)	Kid's Story	0%	63%	37%
	Political Text	0%	31%	69%
	Computer Science Text	0%	56%	44%
	Legal Text	0%	25%	75%
	Poem	0%	44%	56%
	Webpage	6%	75%	19%

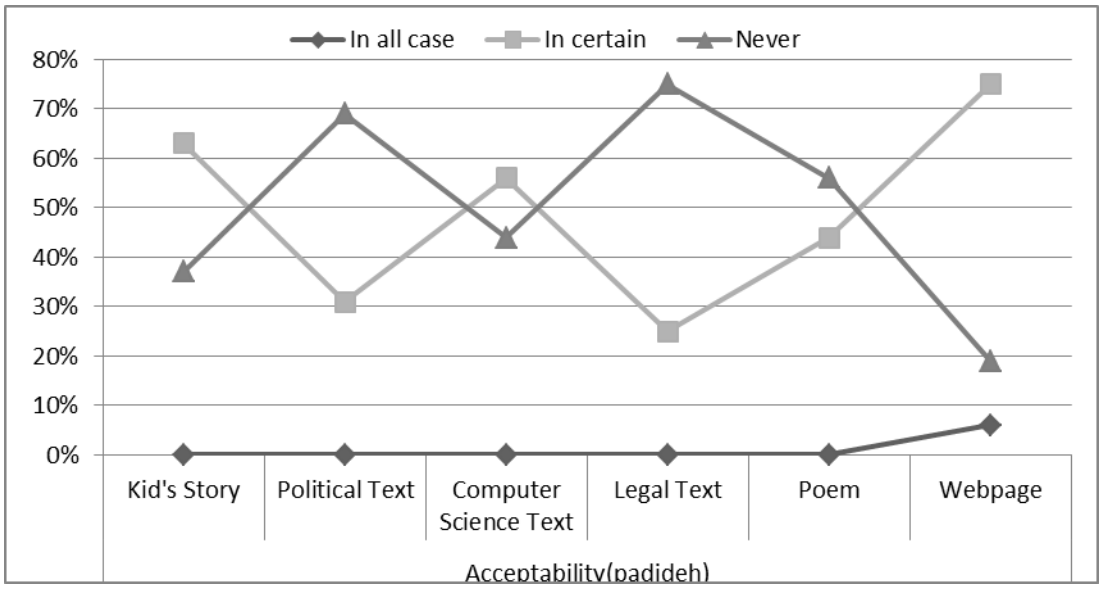

Figure3.4

The answers of the evaluators, in evaluating the Acceptability of six different text types for translations produced by Google Translate, reveals the information, that you can see in Table 3.5 and Figure 3.5.

Table 3.5 Acceptability(Google translate)

Acceptability(Google translate)		In all case	In certain	Never
	Kid's Story	12%	69%	19%
	Political Text	19%	44%	37%
	Computer Science Text	19%	56%	25%
	Legal Text	6%	44%	50%
	Poem	18%	38%	44%
	Webpage	38%	56%	6%

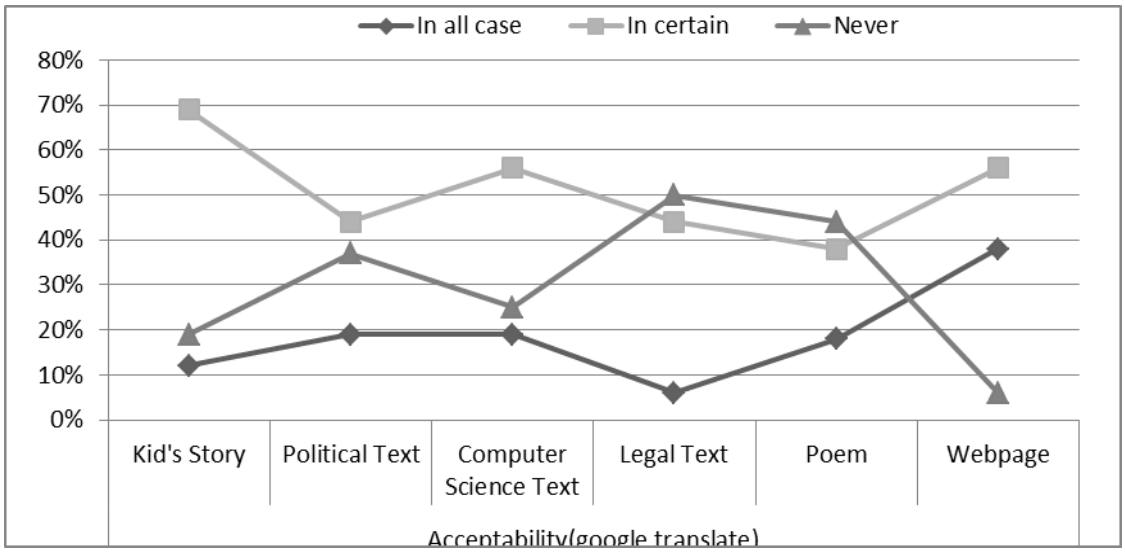

Figure 3.5

The data collected from the answers of sixteen evaluators to the second question in examining the acceptability of the translations produced by Padideh and Google Translate, is showed respectively in Table 3.6, Figure 3.6, and Table 3.7, Figure 3.7.

Table 3.6 Padideh

Padideh		Always	Never	sometimes
	Kid's Story	0%	44%	56%
	Political Text	0%	75%	25%
	Computer Science Text	0%	37%	63%
	Legal Text	0%	44%	56%
	Poem	0%	37%	63%
	Webpage	0%	19%	81%

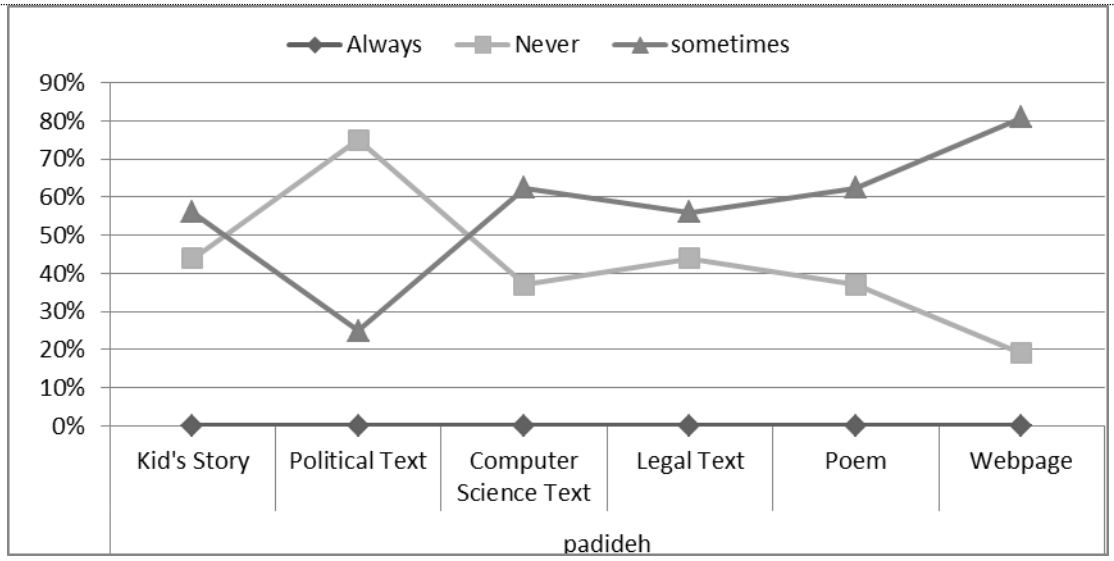

Figure 3.6

Table 3.7 Google translate

		Always	Never	Sometimes
Google translate	Kid's Story	12%	25%	63%
	Political Text	31%	25%	44%
	Computer Science Text	31%	19%	50%
	Legal Text	31%	37%	32%
	Poem	12%	44%	44%
	Webpage	38%	12%	50%

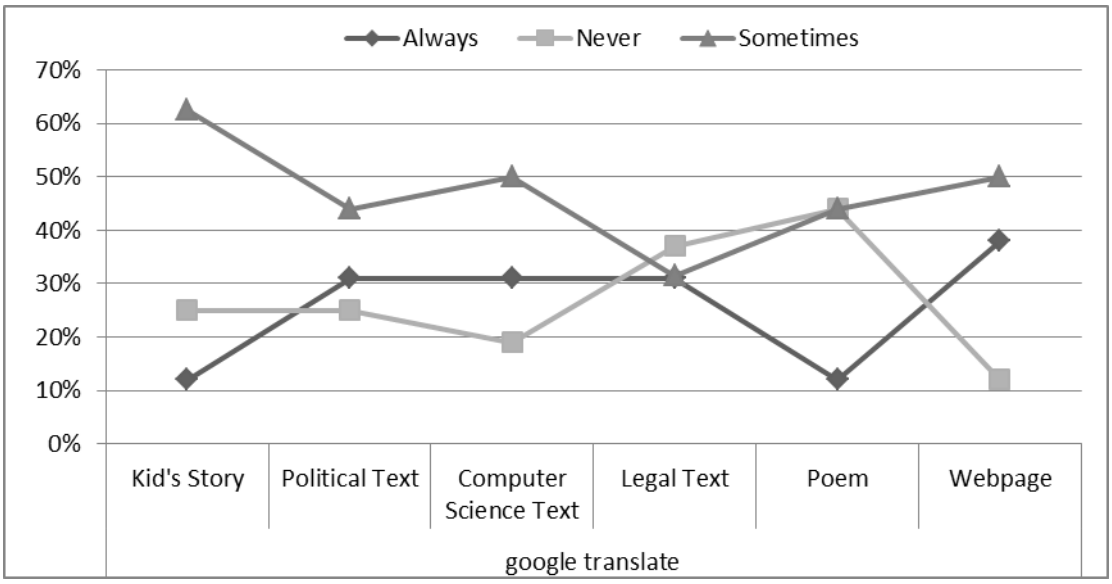

Figure 3.7

3.2.5 Usability

For evaluating the usability of the translations produced by each system, according to Van Slype (1979), evaluators were directly asked about the usability of the systems for each text-type. The information got from analysing the scores rated by sixteen evaluators is showed in Table 3.8 and Figure 3.8.

Table 3.8 Usability(Padideh)

Usability(Padideh)		In all case	In certain	Never
	Kid's Story	0%	75%	25%
	Political Text	0%	37%	63%
	Computer Science Text	6%	69%	25%
	Legal Text	0%	37%	63%
	Poem	0%	63%	37%
	Webpage	12%	88%	0%

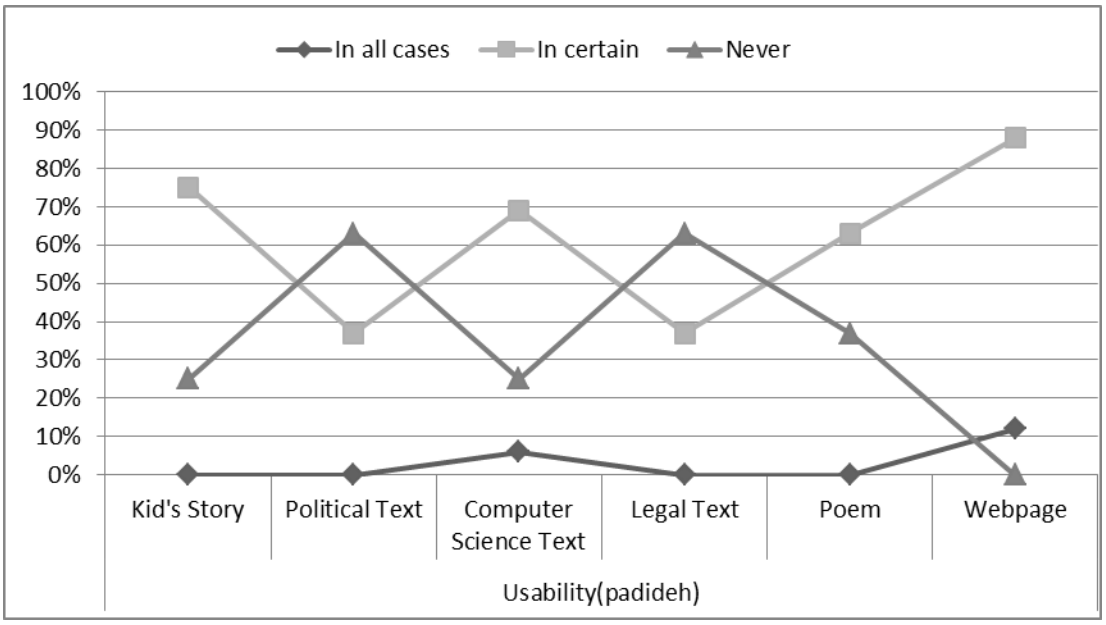

Figure 3.8

Table 3.9 Usability(Google translate)

Usability(Google translate)		In all case	In certain	Never
	Kid's Story	19%	69%	12%
	Political Text	12%	38%	50%
	Computer Science Text	19%	75%	6%
	Legal Text	6%	63%	31%
	Poem	19%	44%	37%
	Webpage	38%	56%	6%

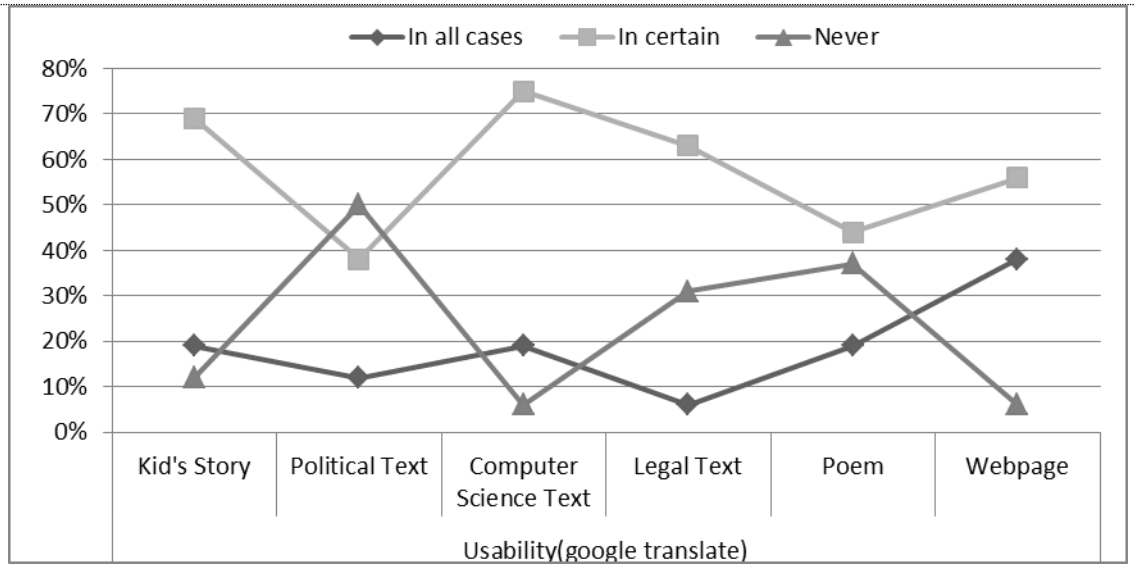

Figure 3.9

3.2.6 Reading Time

According to Van Slype (1979), evaluation method for reading time is, by timing the time spent by the evaluator in reading each text of the sample.

Because of comparative nature of this study, for testing the reading time, the evaluators were asked, for reading which of the translations produced by each system, you have spent much more time?

(The evaluators ought to answer this question for each text-type.)

Collecting data from answers of evaluators in questionnaire revealed that all of the 100% of sixteen evaluators, had spent more time for reading the translations produced by Padideh in relation to Google Translate, for all the six text-types. The evaluators believe that, it was because of incoherency and existence of inaccuracy and mistranslations in translated texts.

4. Conclusion

With all the progress in the field of MT and improvements in machine translation systems, the present evaluation has been carried out and the results thereof have confirmed that there are still serious drawbacks with these systems in translation direction from English to Persian.

In general, both two MT systems under investigation, produced average or below average quality. The total quality performance for each system is as follow:

For Padideh the total average of intelligibility, fidelity and coherence respectively, are: 23.33%, 29.22% and 36.66%. The total average of intelligibility, fidelity and coherence for Google Translate respectively is: 47.77%, 49% and 55%. In fact, their outputs reflect many deficiencies in translating various text types. The end-user can use them to grasp, the general idea of the ST, or translate short, simple texts. Long and complex sentences are especially hard to understand when translated by these systems. One can hardly say that any of them is much better than the others when translating from English into Persian, as much time and efforts are required for intensive post editing of their product.

However, Google Translate is better than Padideh in performing translations of all six text-types, with a quite satisfactory translation. In short, they all need serious improvements.

4.1 Strengths and Weaknesses of the MT Systems Evaluated

4.1.1 Google Translate

Google Translate has obtained higher quality as compared to Padideh software. The TT produced by this system is almost clear especially in translation of computer science text and translation of webpage, it is rather difficult to understand the translation of some text types like literary text, legal text and kid's story, but this is possible after reading the text for two or three times, and the translation produced by this system is rather informative. It also reflects a fairly faithful translation to the ST. In general, the translation of this system is fairly accurate and can be used as a source of information especially in the fields of computer science and translation of webpage. On the other hand, Google Translate has lower reading time, in comparison to Padideh.

4.1.2 Padideh Software

This MT system, which is supposed to have quality level, generally speaking, close to Google Translate, but considering the total system performance, shows that it has given poor results, and it is a lot more below average and has less acceptable quality. Only with certain text types, (e.g. computer science, webpage), although lower than Google Translate, the system can produce translation, where useful information can be extracted. Otherwise, the system performs poorly, compared to Google Translate. The texts are hard to read, some meaning of sentences can be gleaned

with some effort, reflect poor fidelity to the ST and are ill-formed. The general idea is comprehensible, but it is very hard to read large fragments of the MT product. This is attributed to various kinds of mistakes. In short, the translation of this system can only be used by the end user for grasping the general idea of the ST.

4.2 Findings of the Study

To answer RQ1, the data collected during research and analysing them, showed that, the end-users view about machine-generated translation of diverse text-types are as following:

The translation of computer science text is more intelligible than translation of other text-types under investigation in this study, by MT systems. Translation of webpage is the second more intelligible, compared to other text-types. The third rating of intelligibility is kid's story and then translation of political text. The translation of legal text and after that, poem has the lowest intelligibility.

The fidelity of translation of computer science text is the highest, and then with little difference the translation of webpage, the third and forth rating of fidelity belongs to political text and translation of poem, and the last rating of translation fidelity is translation of legal text.

The ranking of coherence rating, are respectively from higher to lower, is: translation of computer science text, poem, kid's story, political text, webpage and legal text.

The translation of kid's story, computer science text and webpage is acceptable for end-users in certain circumstances, and they are willing to use MT, for translation of these text-types, sometimes.

The translation of kid's story, computer science text, poem and webpage, by MT systems, is useful in certain circumstances for the end-users.

The reading time for reading the translations produced by MT systems is high for end-users, because of lack of coherence and mistranslations.

In response to RQ2 we examined both MT systems under investigation, and the results obtained from their comparison are as following:

The total average of intelligibility of translations produced by Padideh and Google Translate, respectively, are: 23.33% and 47.77%.

The total average of fidelity of translations produced by Padideh and Google Translate, respectively are: 29.22% and 49%.

The total average of coherence of translations produced by Padideh and Google Translate, respectively are: 36.66% and 55%.

The acceptability and usability of the translations produced by Google Translate for all the six text-types are higher than translations produced by Padideh, for the same six text-types.

The reading time for reading translation of the six text-types produced by Padideh software is higher than reading time for reading the translations produced by Google Translate for the same text-types.

In summary, considering the obtained results of evaluation parameters, reveals that, the translations produced by Google Translate is more useful and acceptable from the end-users point of view, for all the six text-types.

References

Arnold, D. (1994). *Machine translation: an introductory guide*. Blackwell Pub.

Van Slype, G. (1979). Critical study of methods for evaluating the quality of machine translation. *Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR, 19142*.

Volk, M. (1997, July). Probing the lexicon in evaluating commercial MT systems. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 112-119). Association for Computational Linguistics.