ISSN 2738-2699

YEREVAN STATE
UNIVERSITY

TRANSLATION STUDIES:
THEORY AND PRACTICE

International Scientific Journal

volume 1 | issue 1

YEREVAN STATE UNIVERSITY

Department of Translation Studies

TRANSLATION STUDIES: THEORY AND PRACTICE

International Scientific Journal

volume 1 | issue 1

YEREVAN–2021

METHODOLOGY FOR THE EVALUATION OF MACHINE TRANSLATION QUALITY

ROZA AVAGYAN
ANI ANANYAN

AORIST TRANSLATION AGENCY

Abstract: Along with the development and widespread dissemination of translation by artificial intelligence, it is becoming increasingly important to continuously evaluate and improve its quality and to use it as a tool for the modern translator. In our research, we compared five sentences translated from Armenian into Russian and English by Google Translator, Yandex Translator and two models of the translation system of the Armenian company 'Avromic' to find out how effective these translation systems are when working in Armenian. It was necessary to find out how effective it would be to use them as a translation tool and in the learning process by further editing the translation.
As there is currently no comprehensive and successful method of human metrics for machine translation, we have developed our own evaluation method and criteria by studying the world's most well-known methods of evaluation for automatic translation. We have used the post-editorial distance evaluation criterion as well. In the example of one sentence in the article, we have presented in detail the evaluation process according to the selected and developed criteria. At the end we have presented the results of the research and made appropriate conclusions.

Key words: systems, human metrics, machine translation, methodology

1. Introduction

The aim of the study is to compare the samples translated by Google Translator, Yandex Translator and two models of a translation system designed by Avromic. The comparison is proceeded by our method developed in accordance with a number of international methods and criteria. The main purpose of the comparison is to find out what result these translation programs give when working in Armenian.

Along with the improvement of the quality and widespread translation of artificial intelligence, it is becoming more and more important to continuously evaluate its quality, to work on eliminating the shortcomings, and to use it competently as a tool for modern translators as at present there are no programs that provide quality assurance for the robot translation in compliance with the quality of human translation.

2. Sample, Unit, Sampling

The Sample is the sentences translated by Avromic System and other comparable translation systems (Google, Yandex).

The observed (studied) unit is considered to be the sentence, the title, the content of the table box: a word, a phrase, a sentence.

3. Options

For impartial and proportional distribution, 2% of the units of the document with equal distance from each other were selected for examination, and at least every 50th sentence, but no less than 3 units from each sample.

4. Method Selection and Processing

In the world practice there are various methods for evaluating translation (both human and machine-generated). The evaluation is carried out either by humans or specific evaluation software. To note, automatic and human metrics are different.

On the one hand mechanical evaluation programs (for example, BLEU, METEOR) provide a more objective assessment as they evaluate the translation through formulas, thus excluding the subjective element in the evaluation process. However, on the other hand, mechanical evaluation programs cannot give a complete adequate result, as the aim of the evaluation through those formulas is not to understand the meaning of the text, which leads to an incomplete evaluation.

As currently there are no comprehensive and successful tools which enable to accurately assess automatic translation, while developing our assessment methods and criteria we have considered the most popular and widely used ones.

The most commonly used methodology for evaluating automatic translation is based on the five-point scale of fluency and adequacy. It was developed in 2005 by the Linguistics Data Consortium[1].

Oftentimes the sentence is comprehensible and fluent, but with semantic losses, and vice versa. It is possible to convey the meaning of all the words separately, to translate them all, but the sentence is absolutely incomprehensible.

The first five points are used to estimate how accurate the sentence is, and the other five points are used to assess how well the meaning is conveyed in the translation.

This method also evaluates on a scale of 0-9. However, the disadvantage of this method is that human perception of the text is always subjective, as everyone perceives and understands the meaning of the sentence differently. Also the disadvantage is that the linguistic, grammatical and stylistic errors are not taken into account when evaluating based on these criteria.

The methodology developed by the Defense Advanced Research Projects Agency has been replenished with another criterion: comprehension. The assessment is based on 3 criteria: compliance, fluency and clarity.

In 2007, Chris Callison-Burch et al. suggested including a new method called component-based evaluation in the process of automatic translation evaluation. During this evaluation, the constituent parts of the structure of the original sentence under analysis are selected and the quality of their translation is assessed[2].

The method developed by Michael Denkowski and Alon Lavie provides the classification of errors according to which an error has the greatest impact on the

[1] https://www.ldc.upenn.edu/

[2] http://www.mt-archive.info/ACL-SMT-2007-Callison-Burch.pdf

quality of translation[3]. In this case, there are different types of errors: omission of words, added words, incorrect agreement of words, mispronunciation of words, an incorrectly chosen part of speech, etc.

Makoto Nagao has proposed a 5-point scale for automatic translation evaluation based on linguistic and stylistic analysis[4]. This takes into account whether the meaning of the sentence is clear and perceptible, how well the grammatical rules are observed, whether the correct vocabulary is chosen, the general stylistic conformity is analysed and a score of 1-5 is given. This method is based on subjective perception, therefore the assessment carried out on this scale cannot be complete.

Based on all the above-mentioned methodologies and Nagao's method, we have developed our own method of assessment, aiming to rule out subjective assessment as much as possible. A 10-point scale was selected for evaluation. In this case, only the ideal translation by human and the ideal transfer of the number are valued on a 10-point scale. The following evaluation criteria have been developed to allow more accurate and detailed assessment of the translation quality. The criteria we have chosen are as follows:

1. fluency and clarity,
2. compliance with the original (attention is paid to the omitted, added words),
3. vocabulary compliance with the original (it is estimated that the translation of a word fits the given context),
4. compliance with the grammatical and morphological rules of the given language (the correct sequence of words, the agreement between words, the correct choice of the part of speech are evaluated).

When evaluating automatic translation, it is important to consider the structure (how well the style of the text has been maintained).

The nine most common problems of the style in the translation have been studied on the basis of which our evaluation criteria have been developed:

1. paragraphs and page breaks,
2. line alignment right / left / center / equilateral,
3. maintaining of table and boxes,
4. tabulation,
5. lines,
6. footnotes,
7. forms,
8. images,
9. pointing and numbering.

5. Examination Description

Materials translated by Google Translator, Yandex Translator and Avromic Translation System Model 2 have been examined.

At the same time, the systems designed by Avromic work according to the following principles:

[3] http://www.cs.cmu.edu/~alavie/papers/AMTA-10-Denkowski.pdf

[4] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.8416&rep=rep1&type=pdf

Model I. If the sentence or the phrase does not find an exact match in the system's database (DB), the system separates it into words, searches the DB of the system for each word and in case there is a word that cannot be found in the DB, then that word is sent to the Google Translator with the whole sentence (with both English/Russian and Armenian words).

Model II. If the sentence or the phrase does not find an exact match in the system database, then without being split into words, the system sends it to the Google Translator. Google translates it and the system returns the translated sentence and places it in the text.

The examination was carried out in two stages.

Stage I

In the first stage, the content and semantic correspondence were examined by evaluating the samples on a scale of 0-10 points based the criteria we have developed.

At the same time, the original in Armenian is considered to have 10 points by each criterion. But no translated unit was given 10 points, except for the digitally transferred units, considering only the translation done by human worthy of 10 points.

We have translated one sentence into English and Russian by the translation systems under consideration.

Let's take a closer look at the process of examination through the example of translation quality evaluation.

Russian

Original 1.

Նախատեսվում է ՀՀ ոստիկանության անձնագրային և վիզաների վարչության ստորաբաժանումներում տեղադրել և գործարկել հերթերի կառավարման ավտոմատացված համակարգ, որը բաղկացած կլինի մի շարք ենթահամակարգերից:

Google Translator (code UJ7VD8J8)

Предусматривается установка и запуск системы управления очередями, которая будет состоять из нескольких подсистем в паспортно-визовом отделе полиции РА.

Evaluation

1. The sentence is clear and understandable: 9 points.
2. The word 'ավտոմատացված' is omitted, but it is clear that the system must be automated, so not a significant omission. The word 'վարչության' is omitted: 8 points.
3. The word 'ստորաբաժանումներում' was mistranslated as 'отдел,' whereas it should have been 'подразделения': 8 points.
4. There are no morphological errors, only a single syntactic mistake: in this case the adverbial modifier of place 'в паспортно-визовом отделе полиции РА' would be better to have been placed at the beginning of the sentence: 8 points.

Yandex Translator (code TNFW9U3A)

Планируется паспортов и виз полиции АРМЕНИИ правления подразделениях, установить и запустить в очереди, управления автоматизированная система, которая будет состоять из ряда энтомологии:

Evaluation

1. The sentence is absolutely incomprehensible: 0 points.
2. All the words in the sentence have been translated and there are no omissions or additions: 9 points.
3. The word ‘վարչության’ has been incorrectly translated as ‘правление,’ whereas it should have been ‘управление.’ The word ‘ենթահամակարգեր’ has been incorrectly translated as ‘энтомология,’ instead of ‘подсистемы.’ The word ‘նախատեսվում’ has been incorrectly translated as ‘планируется’ instead of ‘предусматривается.’ The the acronym ‘ՀՀ’ has not been fully translated. The declension of following parts was incorrected in the translated verision: ‘ՀՀ ոստիկանության անձնագրային և վիզաների վարչության ստորաբաժանումներում,’ ‘ավտոմատացված համակարգ,’ so 3 points.
4. Syntactic and morphological rules (e.g. the correct order of the words, the agreement of the members of the sentence) have not been observed: 1 point.

Company. Model I (code P38BMGXY)

Предусматривается установка и запуск автоматизированных систем управления очередями в ряде подсистем в отделах паспортного и визового департамента РА.

Evaluation

1. In general, the sentence is clear and understandable: 9 points.
2. The end of the sentence: “…որը բաղկացած կլինի” has not been translated, which is a significant omission which has caused the meaning to be transferred incorrectly. The word ‘ոստիկանություն’ has not been translated either: 5 points.
3. The word ‘ստորաբաժանում’ has been mistranslated as ‘отдел.’ It should have been ‘позразделение’ instead. The word ‘վարչություն’ has been mistranslated as ‘департамент’ instead of ‘управление.’ The word ‘համակարգ’ is singular in the original, but it has become plural – ‘систем’ in the translation: 6 points.
4. There are no morphological errors, only a single syntactic mistake: in this case the adverbial modifier of place ‘в паспортно-визовом отделе полиции РА’ would be better to have been placed at the beginning of the sentence: 8 points.

Company. Model II (code V3JBQKAY)

Предусматривается установка и запуск системы управления очередями, состоящей из ряда подсистем в паспортно-визовом отделе Полиции Республики Армения.

Evaluation

1. The sentence is clear and understandable: 9 points.
2. The word ‘ավտոմատացված’ is omitted, but it is clear that the system must be automated, not a significant omission. The word ‘department’ is omitted: 8 points.
3. No major vocabulary-related errors seem to exist: 9 points.
4. There are no morphological errors, only a single syntactic mistake. In this case the adverbial modifier of place “в паспортно-визовом отделе Полиции Республики Армения” would be better to have been placed at the beginning of the sentence: 8 points.

English

Original 1.

Նախատեսվում է ՀՀ ոստիկանության անձնագրային և վիզաների վարչության ստորաբաժանումներում տեղադրել և գործարկել հերթերի կառավարման ավտոմատացված համակարգ, որը բաղկացած կլինի մի շարք ենթահամակարգերի:

Google Translator (code UJ7VD8J8)

It is planned to install and launch a queue management system, which will consist of a number of sub-systems in the Passport and Visa Department of the RA Police.

Evaluation

1. The sentence is clear and understandable: 9 points.
2. The word ‘ավտոմատացված’ is omitted, but it is clear that the system must be automated, so it is not a significant omission. The word ‘ստորաբաժանումներում’ is omitted: 7 points.
3. The vocabulary is selected according to the context: 9 points.
4. There are no morphological errors, only a single syntactic mistake: in this case it would have been better to place the adverbial modifier of place in the middle of the sentence, as when placed at the end of the sentence, it causes some misunderstanding. The word ‘sub-systems’ should have been spelled as ‘subsystems’: 8 points.

Yandex Translator (code TNFW9U3A)

It is planned to install and launch an automated system, which will consist of a number of entomologies, in the queue, of passports and visas of the ARMENIAN police Board subdivisions.

Evaluation

1. Many words have been mistranslated, the syntax has been distorted, the text as a whole is incomprehensible, partial translation has been done, words or word combinations have been translated. The detailed analysis is presented in point 4: 3 points.

2. The translation of the word ‘կառավարման’ (management) is omitted: 8 points.

3. Most of the vocabulary is chosen correctly. The word ‘ենթահամակարգեր’ is mistranslated as ‘entomologies,’ instead of ‘units’ which has no contextual relevance to the original. The choice of the word ‘board’ is not correct either. It has the meaning of ‘council,’ but we believe the word ‘department’ would have been more relevant in this case. ‘ՀՀ Ոստիկանությունը’ has been translated as ‘Armenian Police’ (‘Հայոց ոստիկանություն’) instead of ‘the RA Police’: 4 points.

4. There are many grammatical errors as well. The position of the object is wrong. The word ‘անձնագրային’ has been rendered into the target language in the plural form (‘passports’) which is incorrect. There is a wrong choice of adverbial modifier of place: 6 points.

Company. Model I (code P38BMGXY)

It is planned to install and run queues of management automation system, which will consist of a number of sub-systems in RA Police Passport and Visa Department units.

Evaluation

1. The sentence is generally comprehensible: 7 points.

2. There are no significant omissions, the system has recognized almost all units: 8 points.

3. The existing vocabulary is not totally divorced from the context, the word ‘run’ is subject to change though: 8 points.

4. There is a syntactic error. The wrong position of the word ‘queue’ has led to a breakdown of the modifier and the word modified (‘queues of management,’ should be ‘management of queues’). The definite article has been omitted (the RA Police). In this case the adverbial modifier of place would be better to have been placed in the middle of the sentence, as when placed at the end of the sentence it causes some misunderstanding. The word ‘sub-systems’ should have been rendered as ‘subsystems,’ and the ‘automation’ – ‘automated’: 5 points.

Company. Model II (code V3JBQKAY)

It is envisaged to install and launch a queue management system consisting of a number of sub-systems in the passport and visa department of the Police of the Republic of Armenia.

Evaluation

1. The sentence is clear and understandable: 9 points.

2. The word 'ավտոմատացված' is omitted. The word 'ստորաբաժանում' ('unit') is not translated either: 8 points.

3. The vocabulary is selected according to the context: 9 points.

4. In the original, the sentence is complex, while in the translation, it has turned into a simple one. The adverbial modifier of place in this sentence must be put in a middle position, as in this case the meaning of the sentence is misunderstood. The word 'sub-systems' should have been translated as 'subsystems': 6 points.

Stage II

In the second stage, the general format (style) of the document (for example, tables, paragraphs, position) was assessed according to the mentioned criteria. The assessment was carried out on a scale of 0-10 as mentioned above. At the same time, the original Armenian was given 10 points, but no translated version was given more than 9 points, taking into account the element of subjectivity.

The structure was assessed by the following formula:

X = 10 points - 10 x Error Quantity / Quantity in the Original

If (X>9) THEN 9 ELSE X

Average result = Sum of 9 criteria / 9

The number of indicators for each criterion was calculated, the number of errors in each criterion in the set of the given criteria was assessed, and the specific weight of the sum of the given types of errors was subtracted from 10 points. At the same time, if the score was greater than 9 in the formula, then it was given 9 points taking into account the element of subjectivity.

6. Conclusion

We present the separate averaged grades of the content examination of the whole sample, according to the samples, which were also combined and a unified and average grade was provided with the accuracy of 2 digits after the comma.

Translation System	Sample 1			Sample 2			Sample 3			Average rating
	Stage I	Stage II	Average	Stage I	Stage II	Average	Stage I	Stage II	Average	
Google Translator (code UJ7VD8J8)	6,93	7,94	7,43	5,25	6,67	5,96	6,25	6,44	6,35	6,58

Yandex Translator (code TNFW9U3A)	7,95	5,86	6,91	4,82	2,18	3,5	3,33	7	5,17	5,19
Company: Model I (code P38BMGXY)	7,53	8,95	8,24	7,26	7,13	7,2	4	8,44	6,22	7,22
Company: Model II (code V3JBQKAY)	7,95	8,95	8,45	7,26	7,19	7,23	6,17	8,28	7,23	7,65

Table 1. Joint average grades by samples (Russian)

Translation System	Sample 1			Sample 2			Sample 3			Average rating
	Stage I	Stage II	Average	Stage I	Stage II	Average	Stage I	Stage II	Average	
Google Translator (code UJ7VD8J8)	7,25	7,77	7,51	6,77	6,49	6,63	7,5	8,44	7,97	7,37
Yandex Translator (code TNFW9U3A)	8,08	5,86	6,97	6,63	2,3	4,47	5,5	6,89	6,2	5,88
Company: Model I (code P38BMGXY)	8,38	8,95	8,67	8,01	8,16	8,09	6,92	8,56	7,74	8,17
Company: Model II (code V3JBQKAY)	8,43	9	8,71	8,15	7,95	8,05	7,67	8,44	8,06	8,27

Table 2. Joint average grades by samples (English)

As a result of the evaluation through the method mentioned above, we came to the following conclusion: the scores of the English and Russian translations of the studied samples depend on the language and system of translation.

A study of translations done by translators the world over has shown that the Russian ones are the most successful and comprehensible by Yandex Translator, and the English translations are the most successful and comprehensible by Google Translator, which was once again proved by our research findings by calculating the ratings of all translations. Translations of the Company's two translation models are qualitatively superior to those of Yandex and Google translators, and are almost equivalent to each other, but with the first model the terms are better preserved, and in case of the second one it is the overall idea that is conveyed accurately (Tables 1 and 2).

The study also found that the Yandex Translator is the most accurate in terms of appearance (style) and earned a lower score by almost all criteria. The other three systems under study generally show adequate results depending on the criterion. They earned a little bit higher or a little bit lower.

With the advance of modern technologies, the fast pace of life and the time scarcity, it becomes necessary to use translation tools to facilitate and speed up the work of a translation specialist. We believe that the criteria we have developed will enable to offer an accurate and non-subjective assessment of the quality of a translation program and determine whether or not it can be applied to the learning process or translation projects being implemented.

References

Callison-Burch, Chris, Cameron Fordyce, and Philipp Koehn. *Evaluation of Machine Translation*: http://www.mt-archive.info/ACL-SMT-2007-Callison-Burch.pdf.

Denkowski, Michael, and Alon Lavie. *Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks*: https://www.cs.cmu.edu/~alavie/papers/AMTA-10-Denkowski.pdf.

Nagao, Makoto, and Mori Shinsuke. *A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese*: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.84-16&rep=rep1&type=pdf.

Received: 07/07/2020
Revised: 12/03/2021
Accepted: 19/03/2021