International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 15, Issue 2 [February 2026] PP: 101-105

Quality assessment of natural language translators by simulating the telephone game

Dušan Tošić
Faculty of mathematics, University of Belgrade, Serbia

ABSTRACT: *A model TGMn for multiple machine translation of text, using different languages, is described. It is possible estimate the quality of the translated text, also. The inspiration for this model came from the Telephone game. Some situations in which this model can be practically used are described. Concepts related to TGMn are defined and explained using a concrete example. This model can be used to compare available machine translators. Additionally, it can be used to assess the quality of translator for individual natural language without analyzing its implementation, i.e. by treating the translator as a "black box".*
Key words: *natural language, translator, quality, telephone game, assessment*

Date of Submission: 12-02-2026 Date of acceptance: 24-02-2026

I. Introduction

NLP tools (Natural Language Processing) are widely used for translating from one natural language to another. This is a software product that partially or completely replaces a human when translating content from one natural language to another. NLP tools perform so-called Machine Translation (MT) and according to [1], we can distinguish 3 types of machine translation tools: Human Translation with Machine Support, Machine Translation with Human Support, and Fully Automated Translation. Today, the market offers translators from various manufacturers of high quality ([2], [3]), which successfully use artificial intelligence methods. Here, we will only deal with Fully Automated Translation tools when considering the problems with multilingual translation. Various characteristics of these tools can be considered, such as application efficiency, maintainability, expansibility, robustness, and so on. However, the most important characteristic of any translator is the quality of the translation produced by that translator. The quality of the translation depends on: the implementation of the translator, the type of text being translated (in principle, translating a technical manual is easier than translating a poem), as well as, the languages used in the translation. Fully automated translation is prone to mistakes, especially when a single spelling of a word can have multiple meanings (homographs). If a text is automatically translated multiple times using multiple languages, with each severally translation it is possible that some information from the original text may be lost. Can we somehow determine when this loss is significant, or when the resulting translation is poor?

II. Related work

In a large number of papers [4-5], [20-21], the problem of how to evaluate machine translation, that is, how to evaluate the quality of a translator, is discussed. In particular, there are numerous papers in which this issue is considered when translating from English to another language and vice versa ([18-19]) In paper [6], the DeepL translator is described and compared with Google and Microsoft translators. The translation quality control strategy of DeepL is analyzed and its good features are highlighted. The purpose of the paper [7] is to present a comprehensive survey of MTS in general and for English, Hindi and Sanskrit languages in particular. According to [24]: "India has a linguistically rich area - it has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. English is very widely used in the media, commerce, science, technology, and education." If this is taken into account, it is understandable why the majority of MT (Machine Translation) studies relate to Indian languages and English. There are a large number of studies concerning machine translation from English to Hindi languages (such as Tamil, Kannada, Hindi, Telugu) and vice versa ([11-14]). In addition to describing general translation techniques and translation quality analysis, these studies often emphasize the specifics of Indian languages. Due to their particular characteristics, we will not consider them further. The paper [8] reviews the current situation of translation quality assessment at home (North China) and abroad. The paper [9] describes the role of Large language models (LLMs) for their capabilities in natural language processing, particularly generative artificial intelligence (AI). The authors of the paper [10] developed a special methodology for testing the quality of machine translation. They compared five sentences translated from Armenian into Russian and English by Google and Yandex Translator. Two models of the translation system of the Armenian company 'Avromic' are used to find out how effective these translation

systems are when working in Armenian. The paper [15] compares the machine-generated translations produced by Padideh software and Google Translate from English to Persian. Six different text types were used: Kid's Story, Political Text, Computer Science Text, Legal Text, Poem, Webpage to assess the characteristics of the considered translators. The conclusion is, in general, that both MT systems under investigation produced average or below-average quality. A significant number of studies relate to specific texts linked to certain fields ([16-17]) in which, in principle, simpler methods are used to assess quality.

III. Multiple times translation using multiple languages

There are situations in everyday life where there is a need for multiple translations using several languages. Suppose in country X, where the official language is L1, there is a factory that produces a device for which the manual U(L1) is written in L1. Suppose the factory expands production to country Y, where the official language is L2. So, a manual in L2 is needed. The manual U(L2) is generated using an NLP tool. Suppose the same device is produced in the factory plant Y located in province Z, where language L3 is used and a manual U(L3) is required. The manual U(L3) is generated using NLP from U(L2). The question arises whether the translation U(L3) is good enough, i.e. whether U(L3) is practically usable as a manual. To answer this question, it is necessary to compare U(L1) and U(L3) and define when the difference between the original and the translated text is not significant Manuals U(L1) and U(L3) are written in different languages. Comparing the meaning of the two texts is a complicated problem in itself. In order to compare these two manuals, we will translate U(L3) into language L1 (again using an NL-translator) and obtain a modification MU(L1). Now U(L1) and MU(L1) can be compared. We will generalize the problem – instead of dealing with 3 languages, we will work with n (n=2,3,4, …) languages.

IV. Simulating the telephone game

For the purpose of analyzing translations obtained through multiple translations using an NL-translator, we will simulate a game known as the telephone game ([22-23]). In this game, n players (n>2) participate, sitting in a circle or standing in a straight line. Players whisper the phrase (or some text) to their neighbors until it reaches the last player in line. The last player says the phrase (text) out loud and compares it with the phrase (text) announced by the first player (Fig 1).(In Fig. 1 **Pi** represents the i-th player, and **Ti** is the text being communicated.) Also, every player can see their own contribution in changing the initial phrase. (For more details, see [22] and [23].)

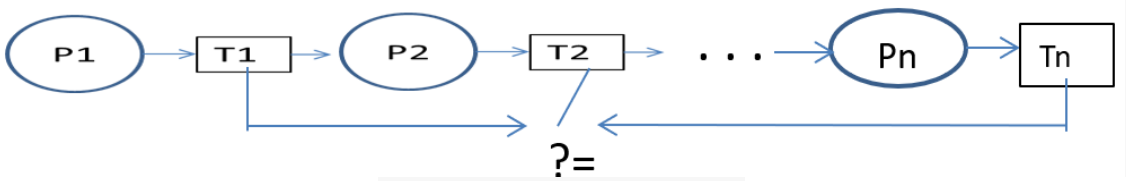

Fig. 1. Shema of telephone game

In the simulation of this game, translators will play the role of players, and the text to be translated (source text) will play the role of the opening phrase (Fig 2). The difference between a telephone game and this simulation appears at the beginning. Namely, in the telephone game the initial phrase is generated by the first player, while in the simulation the source text is immediately known and is communicated (submitted) to the translator (Fig. 2). In one iteration (translation), the text obtained from the previous iteration is translated using the new language. The last in the series of iterations is the one in which the target text CT(Ln) is generated. In order to compare the source text ST(L1) written in language L1 and the target text CT(Ln) written in language Ln (n = 2, 3, ...), it is necessary that these texts are written in the same language. Therefore, we will add another iteration in which CT(Ln) is translated into the language L1. It is an extra iteration that doesn't exist in the phone game. (Fig. 2). It is now possible to compare source and target text using text comparison software tools available on the Internet. In this way, we formed a model (we will call it TGMn where n is the number of iterations in the model).

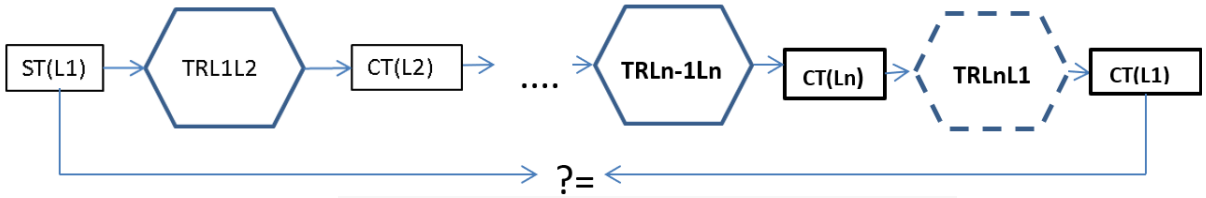

Fig. 2. Simulation telephone game by using translators

In Fig. 2, the following notation is used:

ST(L1) – Source text (to be translated) in language L1
TRL1L2 – translator from language L1 to language L2
CT(L2) – translated source text in language L2
TRLn-1Ln – translator from language Ln-1 to language Ln (n=2,3, ...)
CT(Ln) – target text
CT(L1) – translated target text into language L1 TRLn-1
TLnL1 – translator from language Ln to language L1 (n=2,3,..).

Example 1. If the source text is given in English and needs to be translated into German via French, then L1 = English, L2 = French, L3 = German. Let the source text T(L1) is:

User manuals are vital in every business sector. Basically, they support customers seeking to understand your products and processes. Sometimes, they are even legally required by regulatory bodies.

In this case, CT(L2) is:

I manuâi dal utent a son fondamentâi in ogni setôr comerciâl. In sostance, a supuartin i clients che a cirin di capî i vuês prodots e procès. Cualchi volte, a son ancje domandâts in mût legâl dai cuarps di regolamentazion.

The target text CT(L3) is:

Benutzerhandbücher sind in jedem Geschäftsbereich von entscheidender Bedeutung. Im Wesentlichen unterstützen sie Kunden dabei, Ihre Produkte und Prozesse zu verstehen. Manchmal sind sie sogar gesetzlich von Aufsichtsbehörden vorgeschrieben.

CT(L1) is the target text translated into L1 and reads:

User manuals are crucial in every business sector. Essentially, they help customers understand your products and processes. Sometimes they are even required by regulators.

Here, only Google Translator has been used for translation in all iterations.

V. Using the TGMn model

The TGMn model can be used in various situations to determine the quality of Machine Translations (MTs), without using complex artificial intelligence methods.

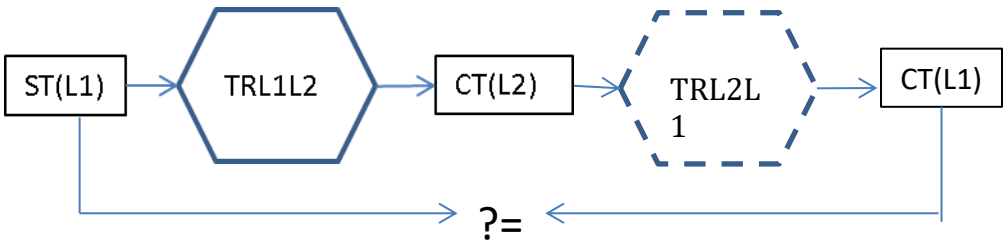

Fig 3. TGM2

When in the model TGNn n=2 (Fig. 3), it is possible evaluate the quality of translation from language L1 ino L2. This possibility could be used in everyday practice. Namely, sometimes it is necessary to translate a text from one language into multiple languages (it is common for a product to have a manual in one booklet written in several languages). Often, manufacturers create a manual for their product in one language, and then translate that manual into k (k=1,2,3,...) languages applying an NL-translator. By using the TGM2 model, it is possible to determine the most suitable translator from the available ones to create a manual in k (k=1,2,3,...) languages.

Example 2. Let the source text be written in English and be as follows:
You can use the Magic Remote Control like a mouse to select and run content on your Smart TV. The Magic Remote Control also supports voice commands and gestures to allow you convenient access to various Smart TV functions.

If we need this text in Serbian, Greek, Romanian, and Hungarian (here k=4), we will translate it into each of these languages, and then translate the target texts into English. Now we can compare the overlap of the source and target text. If the overlap between the source and target text is greater, the translator is better. Here we want to compare 3 translators: Google [25], Microsoft [26], and Yandex [27]. The overlap of source and target texts in percentages can be seen in Table 1. (Thus, we can see that when the given text is translated into Greek using Microsoft Translator, the target text overlaps 82% with the source.

Calculating the average overlap for each translator, we conclude that, in this case, the best results are obtained using the Yundex translator. The results by columns indicate how the translators for individual languages are implemented. By comparing the averages by columns, we can conclude that the best-implemented translator, in this case, is for the Romanian language.

Lang. / TRs	**Ser**	**Gre**	**Rom**	**Hun**	**Average by rows**
Google	**82**	**82**	**84**	**85**	**83,25**
Microsoft	**86**	**82**	**87**	**84**	**84,75**
Yundex	**80**	**83**	**92**	**88**	**85,75**
Average by columns	**82,66667**	**82,33333**	**87,66667**	**85,66667**	

Table 1. Using TGM2 for comparng different translators

Here arises the question of how the comparison between the source and target text is carried out. Comparing texts is a complex task in itself. We will not consider it here. Instead, we will use already existing software tools that can be found on the internet. Some of them are: Cortical.io [29], Twinword's Text Similarity API [28], GoTranscript [30]…

(For the comparison of texts on the basis of which Table 1 was formed, the GoTranscript comparator was used.)

VI. Problems in the application of the TGMn model

There are a number of issues related to the application of the TGMn model. One of the key questions is: what criterion is used to determine whether a translation is acceptable or not? It is difficult to give an answer to this question. Users of translations should also be involved in defining the criteria. Related to TGMn (since some information is lost with each translation), the question arises: after how many iterations (for which n) are the translations poor (unusable)? Here, it is the same problem (how to determine whether a translation is acceptable or not) formulated by model TGMn. Depending on the field of application, the users themselves should determine the criteria. If certain information is lost during each translation, it means that some information may be lost in the last iteration as well. (In the TGMn model, the last iteration is the translation of the target to the source text). Is that loss significant? Based on recent tests, the loss in overlapping source and target text is around 10%. In order to get a more precise answer to this question, a more detailed investigation is needed.

If different translators are used, does a greater difference appear between the target and the source text? Based on a large number of experiments in which different translators and different texts were used, we concluded that the use of different translators (with similar capabilities) does not significantly affect the quality of the translation. For example, if we translate the text from example 2 into Serbian, Greek, Romanian and Hungarian, using only Google translator, the overlap expressed in percentages is shown in row 2 of Table 2. When Google and Microsoft translator are combined, the result is as in row 3 of Table 2.

Lang. / Translator	Serbian	Greek	Romanian	Hungarian
Google-Google	74%	90%	88%	90%
Gogl-Microsoft	77%	92%	93%	92%

Table 1. Using TGM2 for comparing different translators

Another interesting question is whether the result of comparing the source and target text depends on the text comparator? A large number of text comparators can be found on the Internet. They are implemented

differently (some do not consider capitalization and punctuation as mistakes; some of them are adapted for specific texts). Therefore, differences appear in results when using different comparators. It is advisable to conduct testing and choose the most suitable comparator.

VII. Conclusion

By simulating the Telephone Game, a model for testing and evaluating NLP translation was created. Some situations in which this model can be used are described. The model is based on multiple translations of the source text through one or more languages. The quality of the translated text depends on several factors, including the number of translations performed. There are a few issues related to TGMn that need to be addressed for the successful application of this model. Some of them are: which criteria are used to determine whether a translated text obtained using an NL translator is satisfactory? How to compare two texts, i.e., if more of the text comparators are available on the internet, which one is the most suitable? Is it possible to separate a certain class of texts (Political Text, Computer Science Text, Poem,...) and examine which NL translator is the most suitable for each class? Successfully resolving these issues will enable more effective use of the TGMn model.

References

[1]. Shantanoo Dubey: Survey of Machine Translation Techniques, International Journal of Advance Research in Computer Science and Management Studies, Special Issue, Volume 5, Issue 2, pp. 39-51, February 2017.
[2]. Clarriza Heruela: Best Language Translators: How to Pick the Right One, https://www.tomedes.com/translator-hub/best-language-translator, 2025.
[3]. Sujatha R:Top 10 AI Translation Tools for Global Communication, https://www.digitalocean.com/resources/articles/ai-translation-tools, 2025.
[4]. Dušan Tošić, Vesna Tošić: Development of a model for measuring the performance of mashine translators for natural languages, European International Journal of Science and Technology, 10(1), pp. 19-27, 2021.
[5]. A. Way: Quality expectations of machine translation: From principles to practice, DOI:10.1007/978-3-319-91241-7_8, In book: Translation Quality Assessment, pp.159-178, 2018.
[6]. Li Linlin: Artificial Intelligence Translator DeepL Translation Quality Control, https://doi.org/10.1016/j.procs.2024.10.086, Procedia Computer Science, Volume 247, pp. 710-717, 2024.
[7]. Sitender, Seema Bawa, Munish Kumar and Sangeeta: A comprehensive survey on machine translation for English, Hindi and Sanskrit languages, Journal of Ambient Intelligence and Humanized Computing, 14(4), pp. 3441-3474, 2021.
[8]. Leyang Wang & Qingyun Chen: Review of Translation Quality Assessment Research: Current Studies and Development. Sch Int J Linguist Lit, 6(12), pp. 473-477,2023.
[9]. Patteri de Souza, Karen & Koponen, Maarit & Nikolaev, Alexandre, Generative AI for Technical Writing: Comparing Human and LLM Assessments of Generated Content, 2025.
[10]. Ananyan, Ani & Avagyan, Roza: Methodology for the Evaluation of Machine Translation Quality. Translation Studies: Theory and Practice. 1, pp. 124-133, 2021. 10.46991/TSTP/2021.1.1.133.
[11]. Reddy, Mallamma & Hanumanthappa, Prajwal. (2013). Indic Language Machine Translation Tool: English to Kannada/Telugu, 2013, 10.1007/978-81-322-1143-3_4
[12]. Chhabra, Margi Patel: "Issues in Machine Translation of Indian Languages for Information Retrieval", 2021.
[13]. Sanyal, Sugata, and Rajdeep Borgohain. "Machine translation systems in India.", arXiv preprint arXiv:1304.7728, pp. 1-5, 2013.
[14]. Mallamma V. Reddy and Hanumanthappa M.: Semantical and Syntactical Analysis of NLP (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), pp. 3236 – 3238, 2014.
[15]. Haniyeh Sadeghi Azer and Mohammad Bagher Aghayi (Corresponding author): An Evaluation of Output Quality of Machine Translation (Padideh Software vs. Google Translate), Advances in Language and Literary Studies ,Vol. 6 No. 4; pp. 226-237, 2015.
[16]. Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way: Domain-Specific Text Generation for Machine Translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp. 14–30, 2022.
[17]. Ulitkin, Ilya Alekseevich, I. N. Filippova, Natalia Ivanova and Alexey Poroykov. "Automatic evaluation of the quality of machine translation of a scientific text: the results of a five-year-long experiment." E3S Web of Conferences vol. 284, 08001, 2021.
[18]. Chakrawarti, Rajesh & Bansal, Pratosh. Approaches for Improving Hindi to English Machine Translation System, Indian Journal of Science and Technology. 10, pp. 1-8, 2017, 10.17485/ijst/2017/v10i16/111895.
[19]. Pintu Lohar, Maja Popović, and Andy Way: Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy. Association for Computational Linguistics, pp.105-113, 2019.
[20]. Chatzikoumi E. How to evaluate machine translation: A review of automated and human metrics, Natural Language Engineering, Cambridge University Press, pp 1–25, 2019, doi:10.1017/S1351324919000469,
[21]. Leiter, Christoph & Lertvittayakumjorn, Piyawat & Fomicheva, Marina & Zhao, Wei & Gao, Yang & Eger, Steffen, Towards Explainable Evaluation Metrics for Machine Translation. 10.48550/arXiv.2306.13041 pp. 1- 49,2023.
[22]. Telephone game, https://en.wikipedia.org/wiki/Telephone_game.
[23]. https://www.cortical.io/freetools/compare-text/
[24]. Kshirsagar, Suhas: "Machine Translation Projects in India: current status and future prospects", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 289–295, 2014.
[25]. https://translate.google.com/?hl=sr&sl=en&tl=ro&op=translate
[26]. https://www.bing.com/translator
[27]. https://translate.yandex.com/en/dictionary/English-Spanish/translator
[28]. https://www.twinword.com/api/text-similarity.php
[29]. https://www.cortical.io/freetools/compare-text/
[30]. https://gotranscript.com/text-compare