Numerical Assessment of Machine Translation Quality by Method of Near Duplicates Analysis

Kornilov V.S.
South Federal University
Taganrog, Russia
koresh-jr@yandex.ru

Glushan V.M.,
South Federal University
Taganrog, Russia
gluval07@rambler.ru

Lozovoy A.Yu.
South Federal University
Taganrog, Russia
lozovoy@sfedu.ru

***Abstract* — The paper considers modern methods of near duplicates of textanalysisapplied for plagiarism detection, in order to use them in system of machine translation correction. The general concept of the translating machine with automatic correction of inadequacies revealed when comparing the original text and reverse machine translationis presented. Prospects of promotingof the given method are analyzed.**

***Index Terms* — Machine translation, translation quality, translation reversibility, plagiarism search, near duplicate, post-editing.**

I. Introduction

People have been concerned by the problem of text translation quality since the origination of languages and there for translation necessity. Already in the ancient world writers and philosophers, turning to translation problems, pointed out the indispensability of requirements to translation quality assessment [1]. Despiteallthe efforts that have been put forthe lasthalf of a century, the existing methods and techniques of fully automaticmachine translation are far from being perfect. They have a substantial defect, namely, a loss of meaning and incorrect results at translation of the same text from one natural language into another natural language [2]. The translation quality assessment is determined both by the correspondence to the original and"translation equivalence", and from the point of view of the communicative effect achieved by techniques of translation or its pragmatic effect [3]. These approaches are highly subjective; a specific assessment can be formed depending on the goal of ajudge and the intentions of the interpreter. As for machine translation, a numerical evaluation of its quality is decisive for verifying the correctness of the algorithm of the program-translator [1].

In this context, the authors of the article proposed to use a number of existing software tools with the necessary additions for qualitative numerical evaluation of machine translation by means of reverse translation [1] and comparison of the original text and reverse-translated text by method of near duplicates analysis, as the search and analysis of near duplicates is a similar task of intelligent systems in modern information space. This method allows you to compare the syntactic and morphological parameters of the translated text without human input, as well as to analyze and correct the translation algorithm in accordance with the detected inadequacies.

II. Review of the Key Works on the Researching Subject

The possibility of using a reverse translation for translation assessment was first mentioned in the work of Zwilling M.Ya. and Turover G.Ya. "On the criteria for translation assessment" published in 1978 [4], but it did not assume the use of a computer.

The idea to use machine translation facilities for direct and reversetranslation with the subsequent comparison of the original and the reverse- translatedtexts to assess the quality of the translation was set forth in the article "Criteria for the numerical assessment of the quality of a machine-translated text" published earlier by the authors [1].

In that paper, it was proposed to use of the following model: "original text - translation - reverse translation", to achieve the convertibility of the translation, regardless of the translation algorithm itself. Assessment of the translation quality involves comparing the original text with a reverse translation results to obtain a numerical value of the similarity of texts.

In the work [1] it wasalso offered the assessment of translation algorithms by means of the model of a reversible "black box" with the view to their reversibility and the possibility of obtaining a non-distorted reverse translation. As an evaluation measure, the number of corrective operations was assumed until a translation received by the reverse translating machine became identical to the original text.At present, there are many methods for assessmentof machine translation, automated translation, and methods for comparing near duplicates of text. Let us consider some of them:

A. Methods for machine translation assessment:

- the BLEU metric [5], [6] for statistical machine translation, which calculates the ratio of the translated parallel corps to the corresponding phrases in the target language entered into the search engine;
- Amazon Mechanichal Turk [7] (crowdsourcing service from Amazon company), expert systems for the translation assessment, where a large number of anonymous users evaluate the machine-translated text received by different

systems, where at the output a certain normalized comparison coefficient for translation options executed by different systems is achieved;

- assessment in the process of translation and editing of text using standard instructions (assumes the participation of the post-editor). The scheme is presented in Figure 1. This scheme does not involve reverse machine translation, which is a continuation of the idea stated in [3];

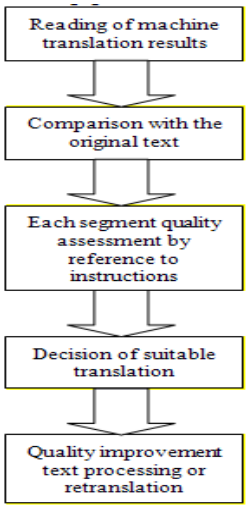

Fig.1. Translation process with post-editor.

B. Methods for comparing near duplicates of unstructured textare developed in the following directions [9]:

a) syntactic methods–period analysis, consisting of symbols, words, word-combinationsor sentences:

- Hamming distance [10];
- Levenshtein distance [11];
- Jaro-Winkler distance [12].

b) lexical methods–hash total detection for words, word-combination, sentences, abstractand further analysis:

- shingles and their modifications [13];
- terms weight coefficient [14]

For implementation of this method, Rule-Based MachineTranslation (RBMT) is suitable. In this type of translation, there are rules for analyzing the source sentence, rules for submitting the presentation at the analysis stage, and, finally, there are rules for the formation of the target sentence from the conveyed material. In the case of direct translation, these rules have a limited scope of application. Rules-based systems are highly accurate and inefficient; in addition, there is a conflict of rules; i.e., when more than one rule becomes applicable to the situation. Thus, the rules should be applied very carefully [15], or the system must be taught (self-taught).

A machine translator must have a set of rules for transforming the text according to the rules of the Waucua triangle [16], also called the Waucua pyramid shown in Figure 2. Professor Bernard Waucua was a translation theorist. Initially trained as a physicist, he became interested in automatic translation when the problem of translation between English and Russian was especially relevant (during the Cold War).

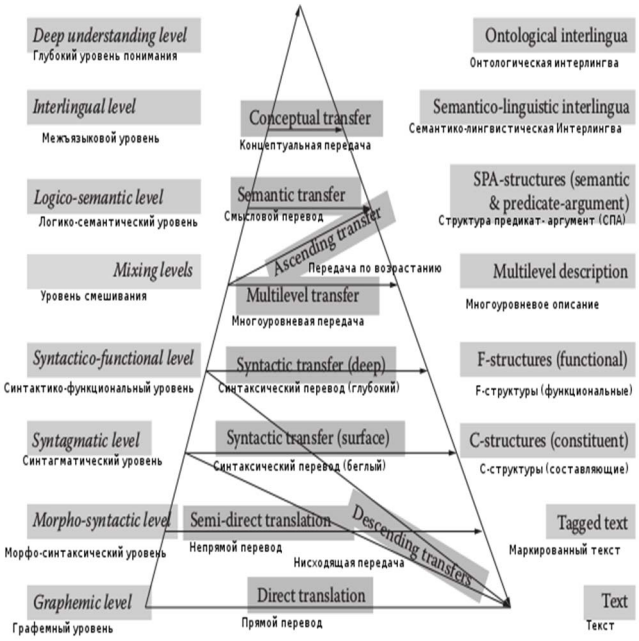

Fig. 2. Waucua Pyramid.

The pyramid shows that translation requires processing at many levels. The left side of the triangle is ascending upward, and the right side is descending. In the left corner, the source language in the right corner is the target language. When we climb up the left side, we conduct an analysis of various types with an input sentence. This processing of the input sentence can include one or more steps or all of the following steps:

1. Morphological analysis;
2. Definition of parts of speech (POS);
3. Identification of the group of the noun and verb (also called shallow analysis or cursory analysis);
4. Parsing with the subsequent introduction of semantics;
5. Discursive output in the form of accompanying links;
6. Pragmatic analysis.

III. Concept of Research

The main concept of the research is to analyze the translation process at all stages of the transformation and establish discrepancies in direct and reverse translation with subsequent adjustments. An example of the implementation of this method is represented in Figure 3.

Information in the text is compared at the grapheme, morpho-syntactic and logical-semantic levels. A model based on recognizing of situations similarities (pattern match search) that involves comparison of the characteristics describing the situation is selected as a model for making decisions on adjustments. This implies the independence of tokens (properties, indicators) [17]. The adjustment process can be performed until the necessary similarity of the original text and the reverse translation is achieved, or until the specified maximum number of operations is achieved, or for a limited time.

The comparison at the grapheme level can be performed with the use of the diff hierarchical algorithm [18], which is based on the calculation of the Hamming distance [19].

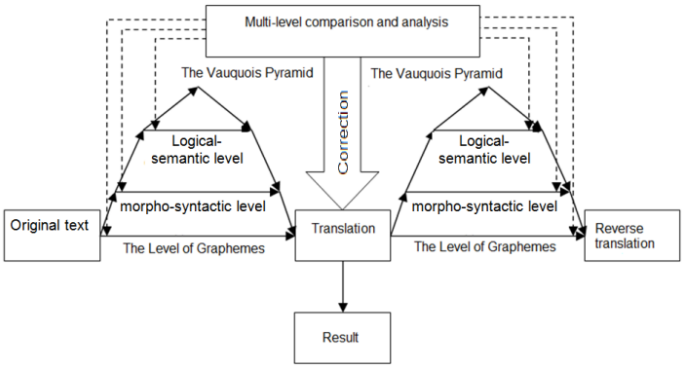

Fig.3. Comparison and adjustment of the machine translation results.

Comparison at the morpho-syntactic level is proposed to be performed using the system of Treeton morpho-syntactic analysis and the Treevial analyzer [20];

Logical-semantic comparison is proposed to be performed using the algorithm for evaluation of students' knowledge in the Intelligent system [21];

IV. Practical Implementation

Practical implementation implies:

- design of a detailed (developed) algorithm of the analysis of near duplicates of machine translation to reveal discrepancies of separate elements of the original text and reverse machine translation;
- design of a detailed (developed) algorithm for making decisions on translation adjustment in case of discrepancies establishing between separate elements of the original text and reverse machine translation;
- code a test application for verification of the method and algorithms using modern online translators as examples;
- design of an online application that complements modern online translators, adjusting the results of their work in an automatic mode on the base of the received discrepancies of the original text and reverse machine translation.

The General scheme of a system of automatic quality assessment and correction of machine translation for a paragraph of text is shown in Fig. 4.

In this system, the human involvement in the translation process is constrained by the choice of the text and set the number of corrective operations.

The scheme was tested on paragraphs of text of technical orientation with a volume of about 1000 characters. The number of corrective operations is 10000. The work of the system makes it possible to improve the quality of translation by approximately 10-15% in the BLEU metric as compared to the first version of the translation.

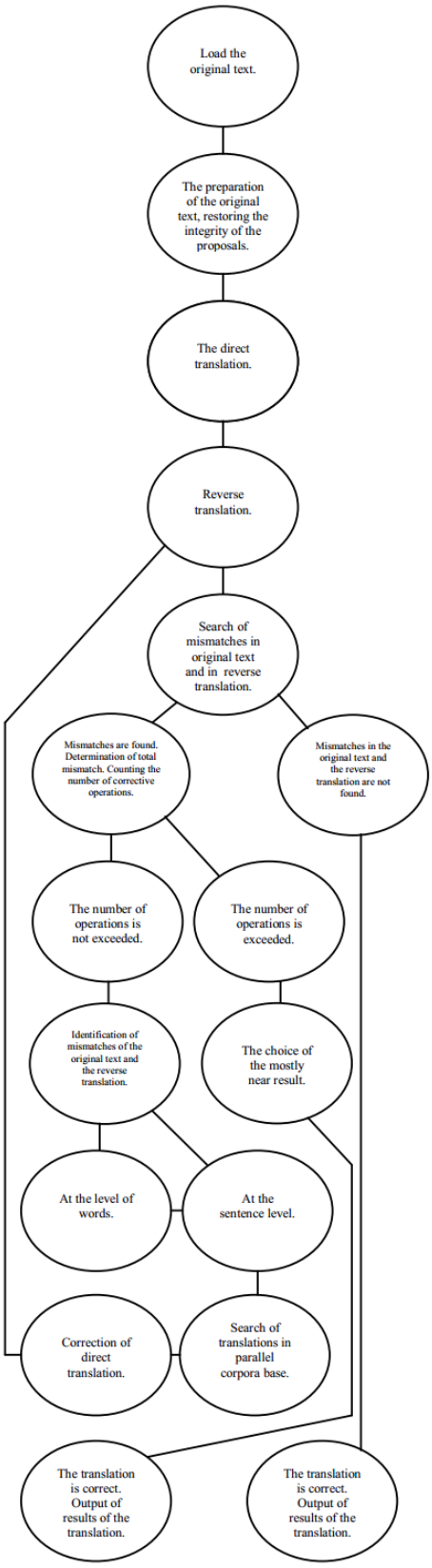

Fig. 4. Automatic quality assessment and correction of machine translations.

V. Conclusion

Modern research in the field of analysis of translated texts for lexical integrity showed that computers successfully cope with simple parts of speech and stable expressions, but allow inconsistencies in cases, figures of speech and the logical and semantic attributes of the sentence when constructing them. Editing of the translated text involves a professional correction of the detected errors or the development of applications for automatic correction of the machine-translated text. In this connection, the technique for improving machine translation is developing in two directions: post-editing machine translation and finalizing the algorithms of the translation itself. Accordingly, the method proposed by the authors for improving the quality of machine translation has no analogues and strikes out a new interdisciplinary direction at the intersection of machine translation and analysis of near duplicates.

It should be noted that machine translation is a highly profitable way of translation ofhuge amounts of text. Automatic translation is used by many translation agencies; consequently, the development of an application for automatic correction of machine translation has serious competitive advantages.

The aim of the authors is to develop an application for automatic correction of machine translation to the level of quality comparable to manual translation.

Summarizing the above, it can be noted from the point of view of common sense, not one translation agency will employ a translator who can translate texts in only one direction, so that a high-quality machine translation system must have the maximum reversibility.

References

[1] Kornilov V.S., Glushan V.M., Criteria for numerical estimation of the quality of the machine-translated text // Information technologies, systems analysis and control. - ITSA-2016; Collected Works of the XIV All-Russian Scientific Conference of Young Scientists, Post-Graduate Students and Students, November 16-19, 2016 - Taganrog: Publishing House of Southern Federal University, 2016 - T.1. - 339 p. P.170-175; Electronic resource: http://rtf.sfedu.ru/ content /itsau_2016_1.pdf; (Date of circulation - 23.02.17);

[2] Liangyou LI, Carla PARRA ESCARTIN, Qun LIU. Combining Translation Memories and Syntax-Based SMT Experiments with Real Industrial Data Special Issue: Proceedings of EAMT 2016 Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 165–177 Electronic resource: http://www.bjmc.lu.lv/fileadmin/ user_upload/lu_portal/projekti/bjmc/Contents/4_2_9_Li.pdf; (Date of circulation – 09.04.17);

[3] Irina Galinskaya, Valentin Gusev, Elena Meshcheryakova, Mariya Shmatova Measuring the Impact of Spelling Errors on the Quality of Machine Translation, Electronic resource: http://www.dialog-21.ru/media/1269/mescheryakovaem.pdf; (Date of circulation - 23.02.17);

[4] Zwilling M. Ya., Tourover G.Ya. About the criteria for assessing the translation // Interpreters' notebooks. №15, M.: IMO, 1978. Electronic resource: http://wt-blog.net/perevodchiku/o-kriterijah-ocenki-perevoda-cvilling-turover.html; (Date of circulation - 01.11.16);

[5] Alexandra Antonova, Alexey Misyurev, building a Web-based parallel corpus and filtering out machine-translated text, Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 136–144, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 24 June 2011; Electronic resource: http://www.aclweb.org/ anthology/W11-1218; (Date of circulation – 10.01.17);

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu; BLEU: a Method for Automatic Evaluation of Machine Translation; Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.; Electronic resource: http://www.aclweb.org/anthology/P02-1040; (Date of circulation – 10.01.17);

[7] Electronic resource: https://www.mturk.com/mturk/ welcome; (Date of circulation - 10.01.17);

[8] Electronic resource: http://blog.perevedem.ru/wp-content/uploads/2014/01/posteditors_2-300x68.png; (Date of circulation - 10.12.16);

[9] Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, Detecting Near-Duplicates for Web Crawling, International World Wide Web Conference 2007, Track: Data Mining Session: Similarity Search, May 8–12, 2007, Banff, Alberta, Canada. pp141-144, Electronic resource: http://www.wwwconference.org/ www2007/papers/paper215.pdf; (Date of circulation - 10.01.17);

[10] Gaborit P., Zemor G., On the Hardness of the Decoding and the Minimum Distance Problems for Rank Codes, Ieee Transactions on Information Theory, Vol: 62 Issue: 12 Pages: 7245-7252;

[11] Khromov N.A., To the problem of identifying near duplicates for detection of plagiarism in scientific publications and reports; Conferences at the Faculty of Physics and Mathematics and Natural Sciences, PFUR, Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems 2012; Peoples' Friendship University of Russia; Electronic resource: http://conf.sci.pfu. edu.ru/index.php/ittmm/2012/paper/view/ 319/0; (Date of circulation - 10.01.17);

[12] D.N. Rubtsov, V.B. Barakhnin, Identification of duplicates in various bibliographic sources, Vestnik NSU. Series: Information technology. 2009. Volume 7, issue 3; Electronic resource: http://www.nsu.ru/ xmlui/ bitstream / handle / nsu / 7131 / 10.pdf; (Date of circulation - 10.01.17);

[13] Zelenkov Yu.G., Segalovich I.V., Comparative analysis of methods for determining near duplicates for Web-documents; Proceedings of the 9th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" - RCDL'2007, Pereslavl-Zalessky, Russia, 2007. Electronic resource: http://rcdl2007. pereslavl.ru/papers/paper_65_v1.pdf; (Date of circulation - 10.01.17);

[14] Zagorulko Yu. A., Salomatin N. V., Seryi A. S., Sidorova E. A., Shestakov V.K. Identification of near duplicates in the automatic formation of thematic collections of documents based on Web publications // Vestn. Novosib. State. University. Series: Information technology. 2013. Vol. 11, no. 4. pp. 59-70. Electronic resource: http://lib.nsu.ru:8081/xmlui/ bitstream/handle/nsu/ 1292 / 06-2013-V11-N4.pdf? Sequence = 1 &isAllowed = y; (Date of circulation - January 12, 17);

[15] A.V. Semenova, V.M. Kurejchik - Review and analysis of the state of the problem of processing textual information in machine translation systems, Informatics, computer science and engineering education. - 2014. - № 2 (17), - Electronic resource: http://digital-mag.tti.sfedu.ru/lib/16/6-2(16)2014.pdf; (Date of circulation - 02.11.16);

[16] Pushpak Bhattacharyya. Machine Translation. CRC Press Taylor & Francis Group. 2015 p.5 Electronic resource: https://play.google.com/store/books/details/ Pushpak_Bhattacharyya_Machine_Translation?id=d4GbBgAAQBAJ; (Date of circulation – 02.11.16);

[17] V.M. Glushan, V.P. Karelin, Use of mathematical models of decision-making in intelligent CAD, Izvestiya SFU. Technical sciences, Thematic issue, Section II. Automation of design; Electronic resource: http://izv-tn.tti.sfedu.ru/wpcontent/uploads/2007/5/12 .pdf; (Date of circulation - 23.02.17);

[18] Gioele Barabucci, Paolo Ciancarini, Angelo Di Iorio, Fabio Vitali Measuring the quality of diff algorithms: a formalization, Computer Standards & Interfaces,Volume 46, May 2016, Pages 52-65 Electronic resource: http://www.sciencedirect.com/ science/article/pii/S0920548915001464; (Date of circulation - 23.02.17);

[19] V.M. Glushan, V.P. Karelin, OL Kuzmenko, Near models and methods of multi-choice in intellectual decision support systems, Izvestiya SFU. Engineering Thematic Issue Section III. Artificial intelligence and near systems; Electronic resource: http://izv-tn.tti.sfedu.ru/wp-content/uploads/2009/4/17.pdf; (Date of circulation - 23.02.17);

[20] A.S. Starostin, M.G. Malkovsky, Algorithm of parsing used in the system of morphosyntactic analysis "treeton", Moscow State University. Of M.V. Lomonosov; Electronic resource: http://starling.rinet .ru/treeton/doc/dialog2007.pdf; (Date of circulation - 23.02.17);

[21] Rajesh Joshi, Satish Kumar, Sunit Kumar, Application of Intuitionistic fuzzy and Hamming Distance Measure in Multiple Attribute Decision Making (MADM), International Journal of Applied Science-Research and Review (IJAS), Electronic resource: https://www.imedpub.com/articles/app lication-of-intuitionistic-fuzzy-and-hamming-distan cemeasure-in-multiple-attribute-decision-making-madm.pdf; (Date of circulation - 23.02.17).