Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy

Dasa Munkova[a], Petr Hajek[b], Michal Munk[a,*], Jan Skalka[a]

[a]*Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia*
[b]*University of Pardubice, Studentska 95, 532 10 Pardubice, Czech Republic*

Abstract

The aim of the paper is to find out whether it is necessary to use all automatic measures of error rate and accuracy when evaluating the quality of machine translation output from the synthetic Slovak language into the analytical English language. We used multiple comparisons for the analysis and visualized the results for each sentence through the icon graphs. Based on the results, we can state that all examined metrics, which are based on textual similarity, except the f-measure, are needed to be included in the MT quality evaluation when analyzing, sentence by sentence, the machine translation output.

Keywords: machine translation; quality; MT evaluation; inflectional language; analytical language; automatic MT measures

* Corresponding author. Tel.: +421-37-6408-672.
E-mail address: mmunk@ukf.sk

1. Introduction and related work

Dynamically changing society and technological progress in translation drew the attention of researchers to examine the importance of translating not only from the aspects of the process (how translators translate) but also product (what is being translated). Incoming technological evolution in translation brings an increase in productivity and consistency of translator's work, broader use of languages, awareness and mutual intercultural communication, but on the other hand, negative perception of translation quality. Translation technologies can be understood as levels of automation of the translation process. Machine Translation (MT) is a system that automatically translates texts from one language to another. Carbonell and Wilks [1] pointed out that MT evaluation is better understood than machine translation. The concept of equivalence in translation is considered to be crucial for both translation and its further evaluation. We follow the conception proposed by House (p.14) [2] in which translation represents the substitution of the original text in source language by the text written in the target language.

Regarding the evaluation of MT quality (especially efficient methods and tools for evaluation of translation quality), experts from the translation industry have deviated from experts from translation studies. Searching for other tools was preferably inspired by time-consumption and ambiguity of adequacy and continuity criteria in machine translation. Papineni et al. [3] note that the methods and tools of manual evaluation are too slow and financially demanding for the development of MT systems; rapid feedback on the quality of translation is important for MT development. Vilar et al. [4] add subjectivity (characteristic for manual evaluation) which can also cause problems in the sense of the evaluator's bias towards machine translation as well as the vague definition of the numeric scale of adequacy and fluency. Snover et al. [5] claim that the inconsistency between the human judgements is reflected in the low correlation between human (manual) assessments metrics. Although criteria for MT quality assessment are still in use, automated evaluation methods have been introduced. They are mostly based on measurements of similarity between MT output (hypothesis) being considered and human translation (reference). Such quality assessment measures are remote from 'human judgment', but they provide a 'quick and dirty' solution that is peculiarly valuable in research and development [6].

Metrics of automatic MT evaluation correlate well with human judgements [3], [5], [7-9] whereby human judgements are in the form of "adequacy and fluency" quantitative scores [10].

Due to globalization, we can see a great ´boom´ of machine translation. Some news agencies, e.g. CNN or BBC, have adopted the MT system to get a 'raw' translation of the news which are originally written in minority languages, to get new information quickly.

Although a number of language researches has been conducted, especially for majority languages such as English; little is known about the MT output of minority language like the Slovak language, however, it belongs to the official languages of European Union.

For the purposes of our study, we will only focus on metrics based on statistical principles and textual similarity, which are widely used in automatic MT evaluation (WER, PER, BLEU, etc.).

The aim of the paper is an MT evaluation, namely the evaluation of MT output generated by Google translate Api. We chose this web application for two reasons. The first reason is its accessibility (free and online), and the second is its multilingualism. Besides, it is the only one that offers translation from/to the Slovak language and also has the largest database of not only of parallel but also comparable texts.

This paper is constructed as follow: section 2 introduces measures of error rate and accuracy for automatic MT evaluation, section 3 describes experiment setting, section 4 presents the experiment results, section 5 consists of discussion and finally section 6 consists of the conclusion and future work.

2. Measures of error rate and accuracy

In this study, similar to studies [11-12], we will focus only on automatic metrics based on textual similarity, i.e. correct word matches between MT output (hypothesis) and reference. Similarity can be mathematically written as a function as follows:

Let *p* be a function that displays each element of the set X into a set of real numbers, i.e. $p : X \times X \rightarrow R$, while:

1. $\forall x, y \in X : p(x, y) \geq 0$ (is a positive number),
2. $p(x, y) = p(y, x)$ (is symmetric),

3. $p(x, y) \leq p(x, x) \Leftrightarrow x = y$.

2.1. Precision, Recall, and f-measure

They are based on the closeness of the hypothesis (MT output) with the reference (human translation), similar to bag-of-words, i.e. regardless of the position of the word in a sentence [13-14].

Precision (*P*) is a measure of how many correct words are present in the hypothesis *h* in regard of reference *r*, i.e. proportion of words in hypothesis *h* (MT output) that are present in reference *r* (human translation):

$$P(h|r) = \frac{|h \cap r|}{|h|}. \quad (1)$$

Recall (*R*) is the number of correct words in MT output (hypothesis *h*) divided by the number of words of reference *r*, i.e. proportion of words in the reference *r* (human translation) that are present in the hypothesis *h* (MT output):

$$R(h|r) = \frac{|h \cap r|}{|r|}. \quad (2)$$

F-measure (F_1) is a harmonic mean of *precision* and *recall*:

$$F_1 = \frac{2PR}{P+R}. \quad (3)$$

2.2. BLEU

BLEU (Bilingual Evaluation Understudy) is a geometric mean of n-gram *precisions* (usually for n-gram of size 1-4) and a *brevity penalty* (*BP*), i.e. length-based penalty to prevent very short sentences as compensation for inappropriate translation [3].

$$BLEU\ (n) = exp \sum_{n=1}^{N} w_n\ log\ p_n \times BP, \text{ where } w_n \text{ is weights for different } p_n, \quad (4)$$

$$BP = \begin{cases} 1, & if\ r > r \\ e^{1-\frac{r}{h}}, & if\ h \leq r \end{cases}, \text{ where } r \text{ is a reference of a hypothesis } h. \quad (5)$$

BLEU represents two features of translation quality- *adequacy* and *fluency* by calculating words or lexical *precision*.

Precision, *Recall*, *F-measure*, and *BLEU-n* are called metrics of accuracy, i.e. the higher the values of these metrics, the better the translation quality. Other widely used evaluation metrics are based on edit distance such as *PER*, *WER*, or *CDER* and are called metrics of error rate, i.e. the higher the values of these metrics, the lower the translation quality.

2.3. WER, PER, and CDER

WER (Word Error Rate) is based on the edit distance (edit operations) and does not allow reordering of words [15]. It calculates the Levenshtein distance between a hypothesis (MT output) and a reference (human translation). The minimum number of edit operations in a sequence *e* (insertions, substitutions, and deletions of words necessary to transform the hypothesis *h* into the reference *r*) is divided by the number of words in the reference.

$$WER(h, r) = \frac{min_{e \in E(h,r)}(I+D+S)}{|r|}, \quad (6)$$

where *r* is a reference of a hypothesis *h*, *I* - insert(*e*), *D* - delete(*e*), *S* - substitute(*e*), and $min_{e \in E(h,r)}$ is a minimal sequence *e* of a set of transformed words *E* (insertions, substitutions, and deletions).

PER (Position-independent Error Rate) is based on *WER* but ignores the ordering of the words in a sentence, i.e. word order is not taken into account [16]. It considers the reference and hypothesis as bags of words and counts the number of times that identical words appear in both sentences (MT output *h* and reference translation *r*):

$$PER = 1 - \frac{|h \cap r| - max(0, h - r)}{|r|}. \quad (7)$$

CDER (Cover Disjoint Error Rate) is a measure oriented towards *recall* but based on the Levenshtein distance [11, 17]. It uses the fact that the number of blocks in a sentence is the same as the number of gaps between them plus one. It requires both, hypothesis and reference to be covered completely and disjointly. Only words in the reference must be covered only once, while in the hypothesis they can be covered zero, one or more times.

$$CDER(h, r) = \frac{min_{e \in E(h,r)}\left(I(e) + D(e) + S(e) + long\ jump(e)\right)}{|r|}, \quad (8)$$

where *insertion* (*e*) – number of adding words, *deletion* (*e*) – number of dropping words, *substitution* (*e*) – number of replacements (in sequence or path *e*), *long jump* (*e*) – number of long jumps, r is reference translation of hypothesis h and $min_{e \in E(h,r)}$ is a minimal sequence e of a set of adding, dropping and replaced words *E*, necessary to transform the MT output *h* into the reference translation *r* [12-13].

3. Experiment

The examined text was original, written in Slovak, consisting of 360 sentences and translated to English by Google Translate API. (Google Translate API (GT) being a free web translation service and is equipped with the largest database of parallel and comparable publicistic texts). GT is also only one online translation service for the Slovak language. The examined dataset is limited because it was obtained from the one-day workshop.

We chose these directions to obtain higher scores of automatic metric BLEU. BLEU metric as a measure for translation quality assessment is not suitable for translation into inflectional languages such as Slovak. English is preferably characterized as an analytical language. Analytical languages do not formally express the difference between nominative and accusative case, i.e. word order in such languages is firmly fixed (SVO) in comparison to Slovak's loose word order. The Slovak language is preferably characterized by synthetic morphology. It is given by numerous forms and morphemes, derivational affixes changing word bases, and modification of words expressing different grammatical categories (e.g. gender, number, case) preferably by one formal feature.

Reference was created by two translators and one native speaker, whose mother tongue is English but also speaks Slovak.

The aim of the research is to evaluate the translation quality of individual sentences of the examined text based on the metrics of automatic MT evaluation from Slovak to English.

To evaluate the quality of the translation, we used our own reliable tool [10, 18]. Its output is a data file represented by 10 variables, where the relative measures of automatic evaluation of translation accuracy ([Precision, Recall, F-measure], [BLEU_1, BLEU_2, BLEU_3, BLEU_4]) and relative measures of automatic evaluation of translation error rate ([PER, WER, CDER]) are calculated for each sentence.

Exploratory techniques will be used to evaluate metrics for automatic MT evaluation. The main thing is to present these data in different ways, to recognize regularities and irregularities, structures, patterns and peculiarities.

For each examined sentence, 10 metrics of automatic MT translation were calculated, which express the quality of the translation in terms of accuracy and error rate. Precision, Recall, F-measure and BLEU measures express translation accuracy, while PER, WER and CDER measures reflect translation error rates.

4. Results

Before the MT evaluation of individual MT sentences itself, we investigated whether it is necessary to include all automatic measures into the analysis of individual MT sentences. We set three global null hypotheses:

a) *There is no statistically significant difference among Precision, Recall, and f-measure*;
b) *There is no statistically significant difference among measures BLEU 1, BLEU 2, BLEU 3, and BLEU 4*;
c) *There is no statistically significant difference among measures PER, WER, and CDER*.

To test the null hypotheses we used adjusted univariate tests for repeated measures, due to violation of the assumption of sphericity for repeated measures ANOVA. The null hypotheses are rejected at the 5% / 0.1% significance level (a) p = 0.02360; b) p = 0.00000; c) p = 0.00000). After rejecting the global null hypotheses, we are interested in among which measures are statistically significant differences.

Table 1. The results of the Tukey HSD test for Precision, Recall, and f-measure.

Measure	Mean	Precision	Recall	F-measure
Precision	65.86		0.00420	0.11508
Recall	61.99	0.00420		0.40383
F-measure	63.50	0.11508	0.40383	

Statistically significant differences (Tab. 1) were identified between Precision and Recall ($p < 0.05$). In contrast, there were no statistically significant differences between the f-measure and the remaining MT measures of accuracy. The strictest measure (Tab. 1) was Recall with an average value of 62% and the least strict was Precision with an average value of 66% for examined MT output.

Table 2. The results of the Tukey HSD test for BLEU 1, BLEU 2, BLEU 3, and BLEU 4.

Measure	Mean	BLEU 1	BLEU 2	BLEU 3	BLEU 4
BLEU 1	63.12		0.00014	0.00014	0.00014
BLEU 2	38.92	0.00014		0.00014	0.00014
BLEU 3	26.21	0.00014	0.00014		0.00016
BLEU 4	18.33	0.00014	0.00014	0.00016	

In the case of BLEU-n (Tab. 2), statistically significant differences were identified among all BLEU-n ($p < 0.05$), where the measure BLEU 4 with an average value of 18% belonged to the strictest and the BLEU 1 to the least strict with an average value of 63% for examined MT output (Tab. 2).

Table 3. The results of the Tukey HSD test for PER, WER, and CDER.

Measure	Mean	PER	WER	CDER
PER	38.05		0.00011	0.00011
WER	53.96	0.00011		0.00036
CDER	47.27	0.00011	0.00036	

Similarly, in the case of MT error rates (Tab. 3), statistically significant differences were identified among all error rates ($p < 0.05$), where the measure WER with an average value of 54% was identified as the strictest one (Tab. 3) and the PER with an average value of 38% as the least strict for examined MT output.

Based on these results, we included all measures into the evaluation of individual sentences of the analyzed MT output. Despite the finding that the evaluation of individual sentences/segments of MT output could be reduced by f-measure, which did not show statistically significant differences with the measures Precision and Recall.

To visualize the multidimensional data, we will use icon graphs, where each icon will correspond to one sentence and will be characterized by the calculated values of the automatic MT evaluation metrics. Icon graphs show the largest and smallest differences in accuracy (Precision, Recall, f-measure, and BLEU-n) and error rate (WER, PER,

and CDER) of MT output (hypothesis) depending on the reference (human translation). The icons representing the translation are displayed in the graph from left to right in the order one by one sentence.

Using these graphs it is possible to identify cases - sentences with the same quality of translation, to divide them into groups, but also to find cases significantly different from the others. By an icon graph, we can visually detect outliers- extreme cases- sentences with significantly different values from the other cases.

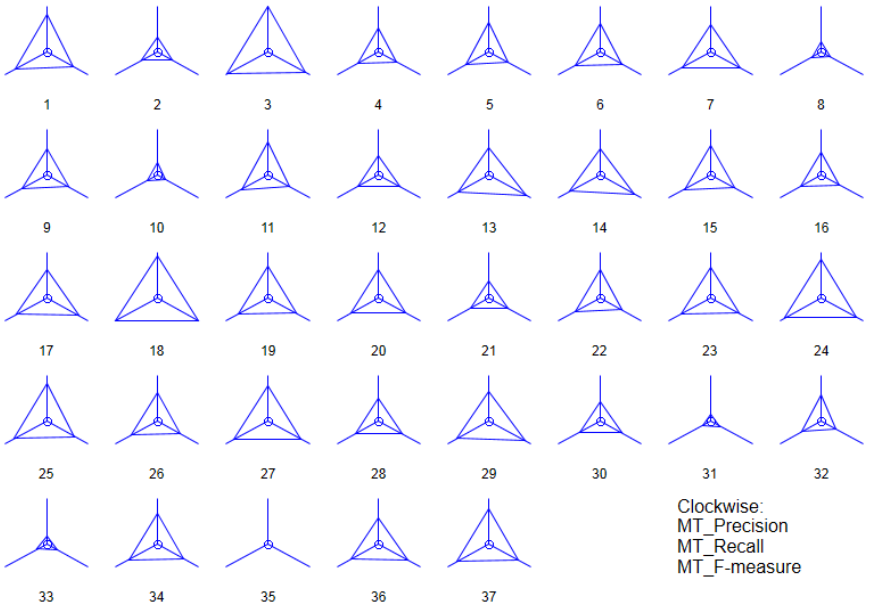

Fig. 1. Values of Precision, Recall, F-measure for first 37 MT sentences; clockwise direction

Based on the automatic metrics of accuracy (Fig. 1), sentences 3, 18, and 24 achieved the highest values in all measures. On the other hand, the smallest values of metrics of accuracy were achieved for sentences 8, 10, 31, 33, and 35. Individual measures of accuracy of MT evaluation (Precision, Recall, and F-measure) are symmetrical in almost all sentences.

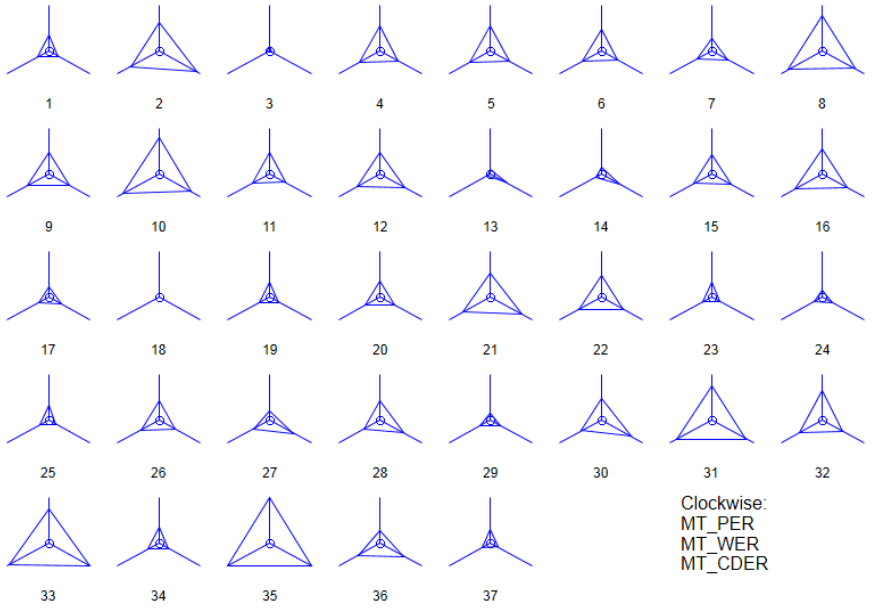

Fig. 2. Values of PER, WER, CDER for first 37 MT sentences; clockwise direction.

Based on the automatic metrics of error rate (Fig. 2), sentences 3 and 18 achieved the lowest values in all measures, i.e. both sentences were translated by GT at the highest translation quality. Vice versa the highest values of metrics of error rate were achieved for sentences 8, 10, 31, 33, and 35. The individual measures the translation error rate of MT evaluation (PER, WER, CDER) show slight asymmetry in some sentences.

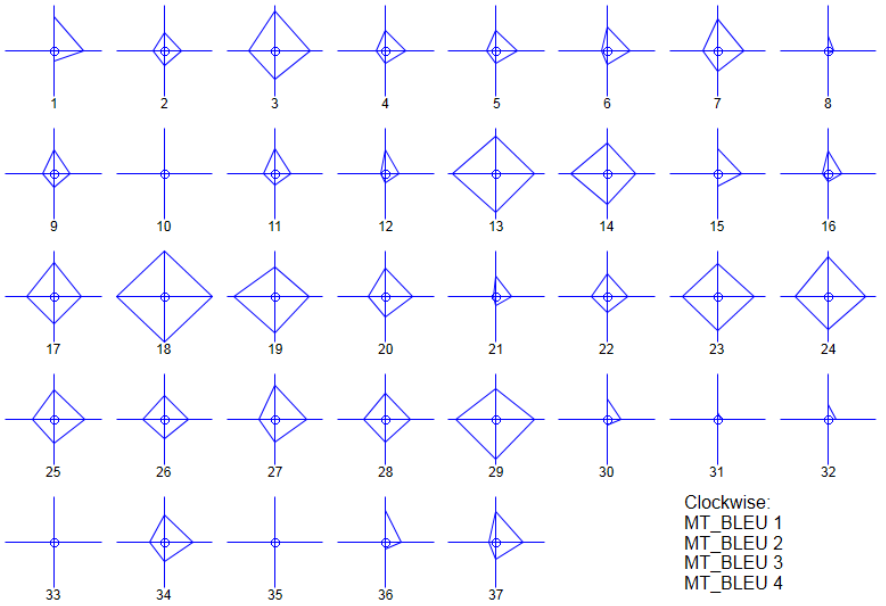

Fig. 3. Values of BLEU-n for first 37 MT sentences; clockwise direction.

Based on the automatic metrics of accuracy (Fig. 3), sentences 13, 18, and 29 achieved the highest values in all measures. On the other hand, the smallest values of metrics of accuracy were achieved for sentences 8, 10, 31, 32, 33, and 35. Individual measures of accuracy of MT evaluation (BLEU_1, BLEU_2, BLEU_3, and BLEU_4) show significant asymmetry in some sentences.

5. Discussion

In terms of accuracy (morphological and lexical) and syntactic structure, sentences 3, 13, 18, 24 and 29 were the best translated, for example:

MT sentence: 18: *The second aspect of commercial interest in machine translation is the high cost of translation and human translation services, especially given the fact that translation is a highly demanding professional job that requires professionalism and enough time*.

MT sentence 13: *In 2012, the Directorate-General for Translation of the European Commission translated 1.76 million pages of legislative documents (primarily legal regulations, documents from/to national parliaments, correspondence with the national governments and others)*. One reason for explaining this result is that the syntax structure in the source sentence is not complicated (without compound/complex sentence). Sentence 18 is an exception, which pleasantly surprised us, but if we look at the sentence in more detail, we find out that the sentence is not so complicated to translate.

On the contrary, the highest error rates were achieved for sentences 35, 33, 31, 8 and 10, for example:

MT sentence 8: *With the erosion of Use multiple languages is also related to some loss of culture, identity of man and total change in his thinking and communicating with the environment.*

MT sentence 35: *We mention two basic, which we discuss in more detail below.*

The results surprised us, for example, the sentence 35 was quite simple in the original (35. We will mention two basic ones that will be discussed in more detail below.); we did not expect a mistake in tenses.

We can claim that the most frequent errors were the ones in subject-predicate relations (identification of predicate, category of tense and mode, errors in the congruence in person, number and gender); numerous errors in syntactic-semantic correlation (mainly in nominal morpho-syntax or verbal morpho-syntax); in lexical semantics (adequate transfer of words, homonymy and polysemy).

After a deeper analysis of the occurred errors, the most common errors were word order (e.g. Machine translation (MT) has become in the last decade significant ...), which is caused by a transfer, i.e. the translation was from an inflectional language with a free word order to an analytical language with a fixed word order. Subsequently, incorrect or omitted words (e.g. most often it was a verb or an article, Slovak does not have articles), synonyms (significant ... important) or abbreviation; we explain it again by a translation error and cultural differences.

Machine translation represents a compromise between quality and quantity, i.e. it is able to translate a large volume of texts in a short time, but of different quality. Besides the language discrepancies arising from different language typology, there are two other challenging factors: language and identity.

6. Conclusion and future work

The paper focused on the translation of the popular-scientific text from the Slovak language into English. The reason why we decided to evaluate the machine translation from the mother tongue into a foreign language was a language system of examined languages. Translation into inflectional languages causes some issues, such as the lower scores for automatic metrics BLEU-n. For the inflectional languages declension, conjugation, inflection, derivation using suffixes and prefixes and, last but not least, loose word order are typical. The examined text consisted of 360 sentences. The longest sentence consisted of 43 words (including prepositions and conjunctions) and the shortest sentence consisted of 6 words. It did not contain specific terminology or too many "long" sentence constructions, abbreviations, or foreign words.

Based on the reference translation, which is considered to be the “gold standard”, we have determined the score of the individual metrics of automatic MT evaluation. Exploratory techniques were used to evaluate metrics for automatic MT evaluation.

Based on our results we can state that it is necessary to include into the evaluation of the quality of machine translation output all measures of error rate and accuracy for each sentence/segment separately. Although the evaluation of the quality of MT output could only be reduced by f-measure, which is understandable since it is a harmonic average of precision and recall. This finding was also confirmed in the analysis of individual sentences, and also the visualization of the individual examined measures through the icon graph has shown (Fig. 1) that f-measure represents only the average value of Precision and Recall.

In future work, we would apply other automatic measures (e.g. Meteor) for the evaluation of MT output from Slovak to English and vice versa.

Acknowledgements

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under contract VEGA-1/0809/18, and the Slovak Research and Development Agency under the contract APVV-18-0473.

This work has been partly developed within the Operational Program: Research and Innovation project “Fake news on the Internet - identification, content analysis, emotions”, co-funded by the European Regional Development Fund.

References

[1] Carbonell, Jaime, and Yorick Wilks (1991) "Machine Translation: An In-Depth Tutorial." *The 29th Annual Meeting of the ACL, Berkeley, California, USA*

[2] House, Juliane (2015) *Translation quality assessment: past and present*. London and NY, Routledge

[3] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, USA*: 311-318.

[4] Vilar, David, Jia Xu, Luis Fernando D’haro, and Hermann Ney (2006) "Error Analysis of Statistical Machine Translation Output." *Language Resources and Evaluation, Genoa, Italy*: 697-702.

[5] Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006) "A study of translation edit rate with targeted human annotation." *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, USA*: 223-231.

[6] Doherty, Stephen (2016) "The Impact of Translation Technologies on the Process and Product of Translation." *International Journal of Communication* **10**: 947-969.

[7] Banerjee, Satanjeev, and Alon Lavie (2005) "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Michigan, USA*: 65-72.

[8] Doddington, George (2002) "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." *Proceedings of the second international conference on Human Language Technology Research, San Diego, California, USA*: 138–145.

[9] Koehn, Philipp (2009) *Statistical Machine Translation.* Cambridge University Press

[10] Munkova, Dasa, and Michal Munk (2014) "An automatic evaluation of machine translation and Slavic languages." *IEEE 8th International Conference on Application of Information and Communication Technologies, Astana, Kazakhstan*: 447-451.

[11] Munk, Michal, Dasa Munkova, and Lubomir Benko (2016) "Identification of Relevant and Redundant Automatic Metrics for MT Evaluation." *MIWAI 2016, Lecture Notes in Artificial Intelligence* **10053**: 141-152.

[12] Munkova, Dasa, and Michal Munk (2015) "Automatic Evaluation of Machine Translation Through the Residual Analysis." *ICIC 2015, Lecture Notes in Artificial Intelligence* **9227**: 481-490.

[13] Munk, Michal, and Dasa Munkova (2018) "Detecting errors in machine translation using residuals and metrics of automatic evaluation." *Journal of Intelligent and Fuzzy Systems* **34** (**5**): 3211-3223.

[14] Munk, Michal, Dasa Munkova, and Lubomir Benko (2018) "Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics." *Journal of Intelligent and Fuzzy Systems* **34** (**5**): 3225-3233.

[15] Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney (2000) "An evaluation tool for machine translation: Fast evaluation for MT research." *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00), Athens, Greece*: 39-45.

[16] Tillmann, Christoph, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf (1997) "Accelerated DP based search for statistical translation." *Fifth European Conference on Speech Communication and Technology, Rhodes, Greece*: 2667–2670.

[17] Leusch, Gregor, Nikola Ueffing, and Hermann Ney (2006) "CDER: Efficient MT Evaluation Using Block Movements.", *11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy*: 241-248.

[18] Munkova, Dasa, and Michal Munk (2016) "Automatic Metrics for Machine Translation Evaluation and Minority Languages." MEDCT 2015, Lecture Notes in Electrical Engineering **381**: 631-636.