Error Classification Using Automatic Measures Based on n-grams and Edit Distance

L'ubomír Benko(✉), Lucia Benkova, Dasa Munkova, Michal Munk, and Danylo Shulzenko

Constantine the Philosopher University in Nitra, 949 01 Nitra, Slovakia
lbenko@ukf.sk

Abstract. Machine translation (MT) evaluation plays an important task in the translation industry. The main issue in evaluating the MT quality is an unclear definition of translation quality. Several methods and techniques for measuring MT quality have been designed. Our study aims at interconnecting manual error classification with automatic metrics of MT evaluation. We attempt to determine the degrees of association between automatic MT metrics and error classes from English into inflectional Slovak. We created a corpus, which consists of English journalistic texts, taken from the British online newspaper The Guardian and their human and machine translations. The MT outputs, produced by Google translate, were manually annotated by three professionals using a categorical framework for error analysis and evaluated using reference proximity through the metrics of automated MT evaluation. The results showed that not all examined automatic metrics based on n-grams or edit distance should be implemented into a model for determining the MT quality. When determining the quality of machine translation in respect to syntactic-semantic correlativeness, it is sufficient to consider only the Recall, BLEU-4 or F-measure, ROUGE-L and NIST (based on n-grams) and the metric CharacTER, which is based on edit distance.

Keywords: Machine translation · Automatic metrics · Error classification

1 Introduction

Machine translation (MT) is one of the most popular natural language processing applications. It is the automatic translation of text from one natural language into another natural language. The quality of the translation, its accuracy or, on the other hand, its error rate, plays a key role in interpersonal communication. Evaluating the MT quality is essential for improving MT systems, as it presents a strong indicator of the correlation between an MT output and its corresponding human translation [1:2]. The biggest issue in evaluating MT quality is an unclear definition of translation quality together with its criteria and measures for translation quality. There are no explicit criteria for "good translation" [1:2]. For this reason, several methods and techniques for measuring translation quality have been designed. In general, they can be divided into a manual approach to MT quality assessment and an automatic approach to MT quality assessment [2]. Both

T. Guarda et al. (Eds.): ARTIIS 2022, CCIS 1675, pp. 345–356, 2022.
https://doi.org/10.1007/978-3-031-20319-0_26

approaches have their advantages, but also their disadvantages. The manual evaluation assesses the translation more likely as a whole, i.e. it assesses cohesiveness and coherence of the translation, but this evaluation is very subjective and time-consuming [3]. The standard criteria used within manual evaluation are fluency (grammatical correctness), adequacy (preservation of the meaning) or usability. In addition to standard criteria of MT quality, human evaluators also use task oriented methods for quality evaluation such as post-editing or error analysis, and/or error classification. Error classification (i.e. identification and classification of errors occurring in a machine translated text) is not only a time-consuming, but also a resource intensive task. It provides a distribution of errors over the defined error classes, but it suffers from low consistency of human evaluators [4].

On the other hand, automatic evaluation brings speed, objectivity, and reusability to the measurement. The objective of automatic MT evaluation is to calculate the numerical score (between 0–1), which represents the quality of MT output and/or the performance of the MT system. This evaluation is less reliable compared to manual evaluation, as the evaluation lies in a lexical comparison of two strings - MT output with reference/human translation - in a target language. Within automatic MT evaluation, there are two main approaches for evaluating quality (MT output) automatically - reference proximity and performance-based techniques [5]. In this study we focus on reference proximity techniques, which are based on statistical principles (lexical similarities) or linguistic features [6]. They compare translation to the human reference in that way, that the closer MT output is to the reference the better the quality is considered to be. Distance between MT output and reference translation is calculated automatically (e.g. WER, TER or CharacTER) or their overlap (e.g. BLEU, F-measure, METEOR or NIST).

Our study aims at interconnecting manual error classification with automatic metrics of MT evaluation. Through error analysis, we point out the degree of association between automatic MT metrics and error classes from English into inflectional Slovak.

The structure of the paper is as follows. The second section introduces automatic MT metrics based on reference proximity. The third section focuses on the methodology of experiment with assumptions, methods, and dataset. The fourth section describes the results of the experiment. Subsequently, the last two sections discuss the obtained results and draw conclusions.

2 Automatic MT Metrics Based on Reference Proximity

Automatic MT metrics provide quantified scores of overall translation quality. They do not require high human effort and they can be used quite easily to compare the performance of two or more MT systems. Therefore, they are not only popular, but also in great demand. Based on their results, MT systems are subsequently developed or optimized.

In this study, we focus on automatic MT metrics that compare MT output with reference based on exact lexical matches between MT words, and/or phrases and reference.

Lexical similarity is a measure of the degree to which the word or phrase of MT output is similar to the corresponding word or phrase in reference. A lexical similarity

of 1 means a total overlap between MT output and reference, whereas 0 means there is no match. *Precision* and *recall* belong to the basic MT metrics [7], where precision is the proportion of words in MT output/hypothesis (Y) that are present in the reference (X), and recall is the proportion of words in reference (X) that are present in the hypothesis (Y). F-measure is a harmonic mean of precision and recall:

$$P = precision(Y|X) = \frac{|X \cap Y|}{|Y|}, \tag{1}$$

$$R = recall(Y|X) = \frac{|X \cap Y|}{|X|}, \tag{2}$$

$$F1 = \frac{2PR}{P+R}. \tag{3}$$

Bilingual Evaluation Understudy (BLEU) is a standard automatic measure, which is a precision-oriented metric. *BLEU-n* [8] is a geometric mean of n-gram precisions with a *brevity penalty* (BP), i.e. penalty to prevent very short sentences:

$$BLEU(n) = exp \sum_{n=1}^{N} w_n \log p_n \times BP \tag{4}$$

where w_n is weights for different p_n,

$$BP = \begin{cases} 1, if\ h > r \\ e^{1-\frac{r}{h}}, if\ h \leq r \end{cases} \tag{5}$$

where r is a reference of a hypothesis h.

The *BLEU* represents two features of translation quality- *adequacy* and *fluency* by calculating words or lexical *precisions* [9]. The *BLEU* score has several variations, depending on the number of words in the reference used to compute the brevity penalty. The IBM version of *BLEU* uses the average value of the length of the reference. The *NIST* version of *BLEU* uses the shortest references to compute the brevity penalty. To not get confused, there exists the *NIST* metric which is not equal to the *NIST* version of *BLEU*, using the arithmetic mean of the n-grams counts instead of the geometric mean, which is used in the ordinary *BLEU-n* metric.

Measure for Evaluation of Translation with Explicit Ordering (METEOR) is a recall-oriented measure. It calculates not only *precision* (like *BLEU*), but also *recall*. Both are combined with a preference to *recall* when calculating the harmonic mean. It is based on a combination of unigram-precision and unigram-recall, and on direct capture of how well-ordered the matched words/phrases in MT outputs are in respect to the reference [10]:

$$METEOR = \frac{10PR}{R+9P}(1 - BP), \tag{6}$$

where the unigram-recall and unigram precision are given by P and R, and

$$BP = 0.5\left(\frac{\#chunks}{\#unigrams\_matched}\right), \tag{7}$$

where chunk (a group of matched unigrams between MT output and reference) is a minimum number of words required to match unigrams in the MT output with corresponding references [11].

NIST [12] is a metric based on *BLEU*. It was designed to improve *BLEU* by rewarding the translation of infrequently used words, i.e. it uses heavier weights for rarer words [11]. The *BLEU* metric calculates n-gram precision with equal weight to each one, but the *NIST* metric calculates how much information is preserved in a particular n-gram.

Character n-gram F-measure (ChrF) is a language- and tokenization-independent metric, which correlates well with human judgments on the system- and segment-level [13]:

$$chrF\beta = (1 + \beta^2)(\frac{chrP \cdot chrR}{\beta^2 chrP + chrR}), \tag{8}$$

where the character n-gram *precision* and *recall* are given by *chrP* (percentage of n-grams in the hypothesis) and *chrR* (percentage of n-grams in the reference). β is a parameter which assigns β times more important to recall than to precision. For instance, if $\beta = 1$, both (precision and recall) have the same weight and if $\beta = 2$, recall is two times more important than precision and vice versa, if $\beta = ½$, precision is two times more important than recall [4, 14].

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) counts the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans [15]. It includes several automatic evaluation measures that determine the similarity between summaries. In this study, we used *ROUGE-N* and *ROUGE-L*. *ROUGE-N* is an n-gram recall between a hypothesis summary and a set of reference summaries. *ROUGE-L* is the longest common subsequence F-measure and counts only in sequence co-occurrences.

The second approach to measure the lexical similarity of two words, and/or phrases is to calculate the minimum edit distance to transform an MT output/hypothesis into a reference (to transform one string into another) through edit operations. Sets of string operations depend on the type of edit distance. One of the simplest sets of edit operations is defined by Levenshtein [16:107-111]:

- Insertion of a character. If $a = uv$, then insert the character x produces uxv. This can also be denoted $\varepsilon \rightarrow x$, using ε to denote the empty string.
- Deletion of a character x changes uxv to $uv(x \rightarrow \varepsilon)$.
- Substitution of a character x for a character $y \neq x$ changes uxv to $uyv(x \rightarrow y)$.

Word Error Rate (WER) counts the Levenshtein distance between the hypothesis and reference, without allowing the words reordering [17]:

$$WER\,(h, r) = \frac{min\#(I+D+S)}{|r|}), \tag{9}$$

where r is a reference of a hypothesis h, I- insertion, D - deletion, and S - substitution.

The minimum number of edit operations (insertions, substitutions, and deletions of the words necessary to transform the hypothesis/MT output into the reference) is divided by the number of words in the reference [7].

Translation Edit Rate (TER) is defined as the minimum number of edit operations required to change a hypothesis/machine translation to an exact match with the reference [18]:

$$TER\,(h, r) = \frac{min\#(I + D + S + shift)}{|r|}), \tag{10}$$

where r is a reference of a hypothesis/machine translation h, I - insertion, D - deletion, S - substitution and shift (number of changes in word order).

CharacTER [19] is an edit distance metric, which is based on character-level and calculates the character-level edit distance while performing the shift edit on word level. Like *TER*, *CharacTER* also calculates the minimum number of character edit operations required to change a hypothesis to the exact match of the reference, divided by the length of the hypothesis:

$$CharacTER\,(h, r) = \frac{min\#(shift + I + D + S)}{|h|} \tag{11}$$

where r is a reference of a hypothesis h, I- insertion, D - deletion, S - substitution and shift (number of changes in word order).

3 Experiment

Our objective is to investigate the relationship between automatic MT metrics and a distribution of errors over the defined error classes. We attempt to determine which of the examined metrics (based on lexical similarity and edit distance) associate the best with individual error classes of a categorical framework for error analysis [20:100]. The examined texts (1903 sentences/3271 segments) were of the journalistic style, taken from the British online newspaper The Guardian. In 2021, the texts were translated by the freely available Neural Google Translate (NGT) engine and subsequently manually annotated by three professionals. The annotation was performed according to the categorical framework for error analysis for translation into Slovak [20:100]. The framework consists of five error classes (categories):

1. Predication,
2. Modal and communication sentence framework,
3. Syntactic-semantic correlativeness,
4. Compound/complex sentences,
5. Lexical semantics.

In this study, we focus only on one particular category - Syntactic-semantic correlativeness - characterizing inflectional languages like Slovak. This category corresponds to the category of *language*, and/or *fluency*, both belonging to the core of harmonized DQF-MQM Error typology [21].

The category of Syntactic-semantic correlativeness is more deeply divided into subcategories: Nominal morphosyntax, Pronominal morphosyntax, Numeral morphosyntax, Verbal morphosyntax, Word order, Other morphosyntactic phenomena, and Others.

3.1 Assumption

Given that the metrics of automatic evaluation are constantly developing, we have been encouraged to examine which of the MT metrics (based on lexical similarity or edit distance) used so far are appropriate and/or best capture the errors that occurred in machine translation into the inflectional language. Besides free word order, inflectional languages are also characterized by inflection and declension. Both linguistic features are particularly captured in the category of Syntactic-semantic correlativeness.

We assume that:

Automatic MT metrics based on lexical similarity (precision, recall, F-measure, ChrF, NIST, ROUGE, METEOR, and BLEU) associate better with the occurrence of errors in a given category than automatic MT metrics based on edit distance (CharacTER, WER, and TER).

To prove our assumption, we used Goodman and Kruskal's gamma. Gamma represents the degree of association between two variables, i.e. the probability of whether two variables are in the same or opposite order.

3.2 Dataset

The dataset consists of machine-translated journalistic texts from English (STs) to Slovak (NMTs). The readability and lexico-grammatical features of our corpus are as follows (Table 1):

Table 1. Dataset composition

Feature type	Feature name	NMTs_SK	STs_EN
Readability	Average sentence length	17.12034	19.26274
	Average word length	5.696361	4.996122
	#short sentences ($n < 10$)	469	395
	#long sentences ($n \geq 10$)	1434	1508
Lexico-grammatical	Frequency of proper nouns	1501	3078
	Frequency of nouns	10070	8627
	Frequency of adjectives	3324	2968
	Frequency of adverbs	933	1667
	Frequency of verbs	5198	6473
	Frequency of pronominals	2371	2124
	Frequency of particles	592	149
	Frequency of foreign words	841	0
	Frequency of interjections	3	3
	Frequency of numerals	617	777
	Frequency of prepositions & conjunctions	6028	6697
	Frequency of interpunction	5958	3547

3.3 Methods

For the metrics as *BLEU*, *NIST*, *METEOR*, and *ChrF* Python Natural Language Toolkit (NLTK) library was used.

```
from nltk.translate.bleu_score import sentence_bleu, sen-
tence_nist, meteor_score, chrf_score
bleu_scores_1.append(sentence_bleu([ref], hyp,
weights=(1,0,0,0)))
bleu_scores_2.append(sentence_bleu([ref], hyp,
weights=(0,1,0,0)))
bleu_scores_3.append(sentence_bleu([ref], hyp,
weights=(0,0,1,0)))
bleu_scores_4.append(sentence_bleu([ref], hyp,
weights=(0,0,0,1)))
nist_scores.append(sentence_nist([ref], hyp, n=1))
meteor_scores.append(meteor_score([ref], hyp))
chrf_scores.append(chrf_score.sentence_chrf(ref, hyp))
```

For *ROUGE*, *TER*, and *WER* open-source libraries were used.

```
import jiwer
import pyter
from rouge_metric import PyRouge
wer_scores.append(jiwer.wer (ref, hyp))
rogue_scores.append(rouge.evaluate_tokenized([hyp], [ref]))
ter_scores.append(pyter.ter(hyp,ref))
```

Precision, *recall*, *F-measure* were implemented separately from the others. The *CharacTER* was implemented as an edit distance function.

4 Results

After manual error classification, we identified 1851 errors in the category of syntactic-semantic correlativeness, of which 394 errors were identified in nominal morphosyntax, 88 errors in pronominal morphosyntax, 4 errors in numeral morphosyntax, 276 errors in verbal morphosyntax, 453 errors in word order, 617 errors in other morphosyntactic phenomena, and 19 errors in the subcategory others.

Based on a Cochran Q test ($N = 3271$, $Q = 1371.86$, $df = 6$, $p < 0.001$) we showed that there are statistically significant differences between the individual subcategories. These results were also proved by *Kendall's Coeff. of concordance* (0.07), where were identified a small agreement, and/or almost no agreement between the examined subcategories.

Based on the results of multiple comparisons, we showed statistically significant differences between Other morphosyntactic phenomena/Word order/Pronominal morphosyntax and other subcategories and, conversely, there were no statistically significant differences between Numeral morphosyntax and Others, or between Nominal morphosyntax and Word order (Table 2).

Table 2. Multiple comparisons: Homogenous groups, $p < 0.05$

	Incidence	1	2	3	4	5
Numeral morphosyntax	0.12%	****				
Others	0.58%	****				
Pronominal morphosyntax	2.69%			****		
Verbal morphosyntax	8.44%				****	
Nominal morphosyntax	12.05%		****			
Word order	13.85%		****			
Other morphosyntactic phenomena	18.86%					****

Using Goodman and Kruskal's gamma, we determined the rank associations between the individual subcategories and the automatic MT metrics based on lexical similarity or edit distance (Tables 3 and 4).

Table 3. Nominal morphosyntax - rank association

Error category & automatic metrics	Valid *N*	*Gamma*	*Z*	*p*-value
Nominal morphosyntax & BLEU-4	3271	**0.08****	2.9868	**0.0028**
Nominal morphosyntax & NIST	3271	**0.05***	2.0931	**0.0363**
Nominal morphosyntax & BLEU-3	3271	0.05	1.9018	0.0572
Nominal morphosyntax & BLEU-2	3271	0.04	1.7635	0.0778
Nominal morphosyntax & precision	3271	0.04	1.5416	0.1232
Nominal morphosyntax & ChrF	3271	0.04	1.4759	0.1400
Nominal morphosyntax & F-measure	3271	0.04	1.4083	0.1591
Nominal morphosyntax & METEOR	3271	0.03	1.2716	0.2035
Nominal morphosyntax & recall	3271	0.03	1.1908	0.2337
Nominal morphosyntax & BLEU-1	3271	0.03	1.1790	0.2384
Nominal morphosyntax & WER	3271	−0.03	−1.1804	0.2379
Nominal morphosyntax & TER	3271	−0.03	−1.2338	0.2173
Nominal morphosyntax & ROUGE1	3271	−0.04	−1.7095	0.0874

(*continued*)

Table 3. (*continued*)

Error category & automatic metrics	Valid *N*	*Gamma*	*Z*	*p*-value
Nominal morphosyntax & ROUGE2	3271	−0.04	−1.7095	0.0874
Nominal morphosyntax & ROUGE-L	3271	**−0.07****	−2.8404	**0.0045**
Nominal morphosyntax & CharacTER	3271	**−0.09*****	−3.6744	**0.0002**

Note: 0.00 to 0.10 (0.00 to −0.10) – trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – very high positive (negative) measure of association; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The subcategory Nominal morphosyntax (Table 3) is partially identified by the metrics *BLEU-4* and *NIST*, where a trivial, but statistically significant degree of positive association was achieved (*Gamma* < 0.1, $p < 0.01/0.05$), similarly, in the case of the metrics *ROUGE-L* and *CharacTER*, there were achieved statistically significant, but trivial degrees of a negative association (*Gamma* < -0.1, $p < 0.01/0.001$).

The automatic metrics *BLEU-4* and *NIST*, both based on precision, associated best with MT errors in the subcategory of nominal morphosyntax. On the other hand, in terms of edit distance, the metric *CharacTER* associated best with this subcategory.

In the case of the subcategories of pronominal morphosyntax and other morphosyntactic phenomena, there were achieved only trivial, statistically insignificant degrees of association between automatic MT metrics and the given subcategories (*Gamma* $\approx$ 0.00).

In the case of the subcategory of numeral morphosyntax, the degree of association oscillates between a low (0.10–0.30 and/or −0.10–−0.30) and a very high (0.70–1.00 and/or −0.70–−1.00) either positive or negative degrees of association (Table 4).

Table 4. Numeral morphosyntax - rank association

Error category & automatic metrics	Valid *N*	*Gamma*	*Z*	*p*-value
Numeral morphosyntax & CharacTER	3271	0.32	1.3681	0.1713
Numeral morphosyntax & TER	3271	0.30	1.2536	0.2100
Numeral morphosyntax & WER	3271	0.30	1.2519	0.2106
Numeral morphosyntax & ROUGE-L	3271	0.30	1.2566	0.2089
Numeral morphosyntax & ROUGE1	3271	0.22	0.9428	0.3458
Numeral morphosyntax & ROUGE2	3271	0.22	0.9428	0.3458
Numeral morphosyntax & BLEU-3	3271	−0.26	−1.0836	0.2786
Numeral morphosyntax & ChrF	3271	−0.40	−1.7115	0.0870
Numeral morphosyntax & BLEU-2	3271	**−0.49***	−2.0522	**0.0401**

(*continued*)

Table 4. (*continued*)

Error category & automatic metrics	Valid *N*	*Gamma*	*Z*	*p*-value
Numeral morphosyntax & METEOR	3271	**−0.57***	−2.4312	**0.0151**
Numeral morphosyntax & BLEU-4	3271	**−0.62***	−2.2520	**0.0243**
Numeral morphosyntax & NIST	3271	**−0.63****	−2.6330	**0.0085**
Numeral morphosyntax & BLEU-1	3271	**−0.65****	−2.7494	**0.0060**
Numeral morphosyntax & precision	3271	**−0.70****	−2.9358	**0.0033**
Numeral morphosyntax & F-measure	3271	**−0.74****	−3.1129	**0.0019**
Numeral morphosyntax & recall	3271	**−0.77****	−3.2238	**0.0013**

Note: 0.00–0.10 (0.00 to −0.10) – trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – very high positive (negative) measure of association; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

In the case of verbal morphosyntax, we achieved similar results as for the subcategory of nominal morphosyntax, i.e. only for *ROUGE-L*, *ROUGE1*, *ROUGE2*, and *CharacTER* were achieved a statistically significant, but trivial degrees of negative association ($Gamma > -0.1$, $p < 0.01$).

We obtained slightly better results for the subcategories Word order and Others, but still with a low positive, and/or negative degree of association. Only for metrics *Recall*, *Precision*, *F-measure*, *BLEU-3*, and *BLEU-4* ($Gamma \geq -0.1$, $p < 0.001/0.01/0.05$) were achieved a statistically significant negative degree of association, in the case of the category of Word order. For the category Others, only the *ChrF* metric has achieved a low, but statistically significant positive degree of association ($Gamma = 0.23$, $p = 0.0345$).

5 Discussion

Metrics like *Precision*, *Recall*, *F-measure*, *BLEU-n*, *NIST*, *METEOR*, *WER*, *TER*, and *ROUGE* are more reliable and have a higher association with linguistic errors within these subcategories: word order, nominal morphosyntax, and numeral morphosyntax. Although they have high associations, the *CharacTER* metric (based on edit distance) has the highest statistical significance among them in nominal morphosyntax. The *ChrF* metric compared to other metrics, which are based on n-grams, showed a poor performance and is not suitable for this linguistic subcategory (error class).

In the case of numeral morphosyntax, the metrics based on n-gram outperform the metrics based on edit distance in all aspects, i.e. in terms of a degree of association with linguistic category, they achieved a higher level of statistical significance ($p < 0.01$). Linguistic categories like verbal morphosyntax, other morphosyntactic phenomena, pronominal morphosyntax, and others do not show the clear associations to automatic metrics (based on n-grams or edit distance) due to approximately the same low degree of association and a low level of statistical significance ($p < 0.05$).

6 Conclusions

The results of our study showed that not all automatic metrics based on n-grams or edit distance should be implemented into a model for determining the MT quality of journalistic texts translated from English into inflectional Slovak. When determining the quality of machine translation in respect to syntactic-semantic correlativeness, it is sufficient to consider only *Recall*, *BLEU-4* or the *F-measure*, *ROUGE-L* and *NIST* (based on n-grams) and the metric *CharacTER*, which is based on edit distance. The results can be also applicable to other inflectional languages.

The results of our study also showed certain pitfalls and limitations that open up space for further research. The first question that arises here is whether automatic MT metrics based on statistical principles (lexical similarity) are suitable for determining the quality of machine translation into the inflectional Slovak language? Or rather to accept into the model automatic MT metrics based on linguistic features? On the other hand, whether the categorical framework used for error analysis is suitable (for translation of journalistic texts from English into Slovak), as the strong associations between automatic MT metrics and the error category under study were not proved.

We consider the size of the corpus to be the main limitation of our study along with the limitation to only one style and genre. In future work, we want to focus on the expansion of our corpus in terms of size and style.

Acknowledgements. This work was supported by the Slovak Research and Development Agency under contract No. APVV-18-0473 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0821/21.

References

1. Chow, J.: Lost in translation: fidelity-focused machine translation evaluation (2019). https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1819-ug-projects/ChowJ-Lost-in-translation-fidelity-focused-machine-translation-evaluation.pdf
2. Castilho, S., Doherty, S., Gaspari, F., Moorkens, J.: Approaches to human and machine translation quality assessment. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. MTTA, vol. 1, pp. 9–38. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_2
3. Sepesy Maučec, M., Donaj, G.: Machine translation and the evaluation of its quality. In: Recent Trends in Computational Intelligence. IntechOpen (2020). https://doi.org/10.5772/intechopen.89063
4. Popović, M.: Error classification and analysis for machine translation quality assessment. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Machine Translation: Technologies and Applications. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_7
5. Babych, B.: Automated MT evaluation metrics and their limitations. In: evista Tradumàtica: Tecnologies De La Traducció, 12 (2014). https://doi.org/10.5565/rev/tradumatica.70
6. Munk, M., Munková, D., Benko, Ľ: Identification of relevant and redundant automatic metrics for MT evaluation. In: Sombattheera, C., Stolzenburg, F., Lin, F., Nayak, A. (eds.) MIWAI 2016. LNCS (LNAI), vol. 10053, pp. 141–152. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49397-8_12

7. Munk, M., Munkova, D.: Detecting errors in machine translation using residuals and metrics of automatic evaluation. J. Intell. Fuzzy Syst. **34**, 3211–3223 (2018). https://doi.org/10.3233/JIFS-169504
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia (2002)
9. Munk, M., Munkova, D., Benko, L.: Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics. J. Intell. Fuzzy Syst. **34**, 3225–3233 (2018). https://doi.org/10.3233/JIFS-169505
10. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (ACL-05), pp. 65–72. Michigan (2005)
11. Wołk, K., Koržinek, D.: Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking (2016)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, pp. 138–145 (2002)
13. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Stroudsburg, PA, USA (2015). https://doi.org/10.18653/v1/W15-3049
14. Popović, M.: chrF deconstructed: beta parameters and n-gram weights. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 499–504. Association for Computational Linguistics, Stroudsburg, PA, USA (2016). https://doi.org/10.18653/v1/W16-2341
15. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
16. Jurafsky, D., Martin, J.: Speech and Language Processing (2020)
17. Nießen, S., Och, F.J., Leusch, G., Ney, H.: An evaluation tool for machine translation: Fast evaluation for MT research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), pp. 39–45 (2000)
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
19. Wang, W., Peter, J.-T., Rosendahl, H., Ney, H.: CharacTer: translation edit rate on character level. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp. 505–510. Association for Computational Linguistics, Stroudsburg, PA, USA (2016). https://doi.org/10.18653/v1/W16-2342
20. Vaňko, J.: Kategoriálny rámec pre analýzu chýb strojového prekladu. In: Munkova, D. and Vaňko, J. (eds.) Mýliť sa je ľudské (ale aj strojové), pp. 83–100. UKF v Nitre, Nitra (2017)
21. Lommel, A.: Metrics for translation quality assessment: a case for standardising error typologies. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. MTTA, vol. 1, pp. 109–127. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_6