applied sciences

MDPI

Article

Evaluation of English–Slovak Neural and Statistical Machine Translation

Lucia Benkova [1,*], Dasa Munkova [2], Ľubomír Benko [1] and Michal Munk [1,3]

[1] Department of Informatics, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, SK-949 01 Nitra, Slovakia; lbenko@ukf.sk (Ľ.B.); mmunk@ukf.sk (M.M.)
[2] Department of Translation Studies, Constantine the Philosopher University in Nitra, Štefánikova 67, SK-949 74 Nitra, Slovakia; dmunkova@ukf.sk
[3] Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, CZ-532 10 Pardubice, Czech Republic
* Correspondence: lucia.benkova@ukf.sk

Abstract: This study is focused on the comparison of phrase-based statistical machine translation (SMT) systems and neural machine translation (NMT) systems using automatic metrics for translation quality evaluation for the language pair of English and Slovak. As the statistical approach is the predecessor of neural machine translation, it was assumed that the neural network approach would generate results with a better quality. An experiment was performed using residuals to compare the scores of automatic metrics of the accuracy (BLEU_n) of the statistical machine translation with those of the neural machine translation. The results showed that the assumption of better neural machine translation quality regardless of the system used was confirmed. There were statistically significant differences between the SMT and NMT in favor of the NMT based on all BLEU_n scores. The neural machine translation achieved a better quality of translation of journalistic texts from English into Slovak, regardless of if it was a system trained on general texts, such as Google Translate, or specific ones, such as the European Commission's (EC's) tool, which was trained on a specific-domain.

Keywords: neural machine translation; statistical machine translation; text analysis; automatic evaluation; Slovak language; English language

Citation: Benkova, L.; Munkova, D.; Benko, Ľ.; Munk, M. Evaluation of English–Slovak Neural and Statistical Machine Translation. *Appl. Sci.* **2021**, *11*, 2948. https://doi.org/10.3390/app11072948

Academic Editor: Julian Szymanski

Received: 9 March 2021
Accepted: 22 March 2021
Published: 25 March 2021

1. Introduction

Machine translation (MT) is a sub-field of computational linguistics that primarily focuses on automatic translation from one natural language into another natural language without any intervention [1].

Neural machine translation (NMT) is an approach that is used by many online translation services, such as Google Translate, Bing, Systran, and eTranslation. NMT works on the principle of predicting the probability of a word sequence. Using sequence-to-sequence models, NMT took a huge leap forward in accuracy [1]. It uses a deep neural network to process huge amounts of data, and is primarily dependent on training data, from which it learns. If there is a substantial dataset for training the model, then NMT can process any language pair, including languages that are difficult to understand. NMT allows the processing of text in various language styles, such as formal, medical, financial, and journalistic. Neural networks are flexible and can be adjusted to many needs. This is ensured by several parameters with weights and interconnections between individual nodes of the neural network. In the field of natural language processing (NLP), this indicates the ability to create millions of language pairs. Word representation in NMT is simpler than in phrase-based MT systems. Each word is encoded into a vector of a unique size and direction. The language and translation model is replaced by a single-sequence model, which gradually forms word by word. To predict the probability of a word sequence, it is necessary to use a neural network that can remember the previous sequence of words in a

sentence. Feedforward neural networks (FNNs) process inputs independently from the rest of the sentence. Recurrent neural networks (RNNs) are ideal for this problem. RNNs are used for both the encoder and decoder, which are the basis for machine translation using NMT. The encoder is used to process the source sentence, which is read and encoded into a vector that captures the "meaning" of the input sequence. The decoder processes this vector to produce an output sequence in the target language.

Google Translate (neural machine translation) is one of the most widely used online translator systems today, thanks to its more natural translations and its ability to process so-called "zero-frame translations". Natural translation provides training in algorithms on large amounts of data. The ability of the so-called "zero-range translation", i.e., direct translation from the source language to the target language without the need for an intermediate step-translation into English, is an improvement over the previous version of Google Translate (statistical machine translation or rule-based machine translation).

American linguist Noam Chomsky [2] argues that language learning is an innate ability that is typical only for humans due to the neural structure of the human brain, and that the environment only shapes the contours of this network into a particular language. Thanks to the so-called "universal grammar", i.e., a set of syntactic rules and principles that are the same for all languages, it is possible to learn to create sentences and interpret them in any language. Christensen [3] summarized the arguments for and against the usage of universal grammar in his article. Christensen's main argument for the existence of universal grammar concludes that even though children do not have the required linguistic input to "learn" grammatical structures, they still make virtually no mistakes. The use of universal grammar could probably lead to more adequate outputs of machine translations. It would be most suitable for low-resource languages and languages with inflected morphologies and loose word orders. Despite that, the debate about universal grammar is still ongoing, with many supporters on both sides.

The need to evaluate MT systems through their outputs (products) is not a current phenomenon; it has existed as long as machine translation itself [4]. In general, we can approach quality assessment manually or automatically. Manual evaluation is one of the most desired, but it is often referred to as subjective and time- and labor-consuming. In addition, the different sensitivities to errors of the evaluators and their different skills cause their inconsistency in evaluating the quality of machine translations [5]. The second approach to quality assessment, which should solve the above-mentioned shortcomings of manual evaluation, is automatic. Metrics of automatic evaluation may offer a valuable alternative to manual evaluation [4]. Automatic evaluation metrics are used to track developmental changes in one particular MT system, but can also be used to measure differences in quality between two different MT systems [6]. They compare the MT output (candidate/hypothesis) to one or several references (human translations). They measure the closeness and/or compute the score of their lexical concordance. According to the criterion of lexical concordance, automatic metrics of MT evaluation can be divided into metrics of accuracy (matched n-grams) and metrics of error rate (edit distance) [7].

Research Objectives

This research aims to compare the outputs of a phrase-based statistical MT system with the outputs of a neural MT system. The comparison will be linguistically motivated and conducted based on automatic translation quality metrics (BLEU_1, BLEU_2, BLEU_3, and BLEU_4). We will also distinguish between two MT systems in both their SMT and NMT versions (Google Translate—GT_SMT and GT_NMT—and the European Commission's MT tools—mt@ec and eTranslation). The aim is comprised of two partial objectives.

The first objective deals with Google Translate (GT); i.e., based on BLEU_n metrics, we compare the quality of statistical machine translations and neural machine translations of journalistic texts from English into Slovak.

The second objective is similar to the first, but it deals with the European Commission's MT tool; i.e., based on BLEU_n metrics, we compare the quality of statistical machine translations and neural machine translations of journalistic texts from English into Slovak.

We assume that the machine translation system based on neural networks will achieve a better quality (accuracy) than the statistical machine translation system, regardless of whether it is an MT system trained on general texts (Google Translate) or trained on domain-specific parallel texts (the European Commission's MT tool).

The structure of the paper is as follows: We provide a brief review of related work in the field of MT focused on SMT and NMT (Section 2). Then, we describe the applied research methodology and dataset (Section 3). The Section 4 provides research results based on an automatic MT evaluation and a residual analysis (Section 4). The Section 5 offers a discussion of the results, and the Section 6 summarizes the contributions and limitations of the presented research.

2. Related Work

NMT is a model based on an encoder–decoder framework that includes two neural networks for an encoder and a decoder [8,9]. The encoder maps the source text input token into a vector and then encodes the sequence of vectors into semantically distributed semantic representations that serve for the decoder to generate the target language sentence [10]. Based on a field review done by [10], an encoder–decoder based on a recurrent neural network [8,11] was changed through the convolutional neural network [9] into a self-attention-based neural network transformer [12]. This transformer is state of the art in terms of both quality and efficiency. Within the transformer, the encoder consists of N identical layers, which are composed of two sublayers: a self-attention sublayer and a feedforward sublayer. The output is the source-side semantic representation. The decoder has also N identical layers composed of three sublayers [10]:

- Masked self-attention mechanism summarizing the partial prediction history;
- Encoder–decoder attention sublayer determining the dynamic source-side contexts for current prediction;
- Feedforward sublayer.

In this Section, we focus on papers dealing with a linguistic evaluation of the MT outputs of various MT systems. Biesialska et al. [13] analyzed the performance of the statistical and neural approaches to machine translation. They compared phrase-based and neural-based MT systems and their combination. The examined language pairs were Czech–Polish and Spanish–Portuguese, and the authors used a large sample of parallel training data (they used a monolingual corpus and a pseudo-corpus). The authors applied back translation into their MT system and observed the results using the BLEU (Bilingual Evaluation Understudy) score [14]. The results showed that for the Czech–Polish language pair, the BLEU score was relatively low, which was explained by the language distance.

Webster et al. [15] focused on examining the applicability of NMT in the field of literary translation. The authors studied the outputs of the Google Translate and DeepL MT systems. The dataset consisted of four English classic novels and their translations into Dutch. The authors compared the human translations with two MT outputs. The error analysis was done using the SCATE error taxonomy [16,17], which is a hierarchical error taxonomy based on the distinction between accuracy and fluency errors. The results showed that most of the sentences contained errors. More fluency errors were identified than accuracy errors. The literary NMT had more difficulty in producing correct sentences than accurately representing what was being said in the source text.

Yu et al. [18] proposed a robust unsupervised neural machine translation with adversarial attack and regularization on representations in their paper. The authors encouraged the encoder to map sentences of the same semantics in different languages to similar representations in order to be more robust to synthesized or noisy sentences. The authors used the BLEU score to evaluate both models for the English–French language pair, which was used as the validation set. The results showed that the model with epsilon correction had a more stable training curve and a slightly better translation score.

Haque et al. [19] focused on the investigation of term translation in NMT and phrase-based SMT. They created their own phrase-based SMT systems using the Moses toolkit. On the other hand, the NMT systems were created using the MarianNMT [20] toolkit. The parallel corpus was the English–Hindi language pair, and it served to create the baseline transformer models. An evaluation of the systems was conducted using the following automatic evaluation metrics: BLEU, METEOR [21], and TER (Translation Error Rate) [22]. The results show that the English to Hindi translation produced reasonable BLEU scores for the free-word-order language of Hindi, with a better score for the NMT. In the case of Hindi to English translation, moderate BLEU scores were produced. Statistically significant differences were identified for all evaluation metrics between the phrase-based SMT and NMT systems. In addition, due to the unavailability of a gold standard for evaluation of terminology translations, the authors created an approach that semi-automatically created a gold-standard test set from an English–Hindi parallel corpus. The sentences for the test were translated using their phrase-based SMT and NMT systems.

Dashtipour et al. [23] focused on a sentiment analysis approach based on a novel hybrid framework for concept-level analysis. It integrated linguistic rules and deep learning to optimize polarity detection. The authors compared the novel framework with state-of-the-art approaches, such as support vector machine, logistic regression, and deep neural network classifiers. A sentiment analysis for Persian in which the current approaches were focused on word co-occurrence frequencies was examined. The proposed hybrid framework achieved the best performance in comparison to the other sentiment analysis methods mentioned.

Almahasees [24] compared the two most popular machine translation systems, Google Translate and the Microsoft Bing translator. Both systems used statistical machine translation. The examined language pair was English and Arabic. The dataset consisted of sentences extracted from a political news agency. A comparison of the MT outputs was conducted using the BLEU automatic evaluation metric. The results were in favor of Google Translate; Bing generated different sentences. The limitation of the experiment was that its corpus consisted of 25 sentences.

Almahasees [25] compared two machine translation systems: Google Translate and the Microsoft Bing translator. The dataset consisted of journalistic texts written in Arabic, and the target language was English for the input texts. Both MT systems used NMT, and they were compared using linguistic error analysis and error classification. The author focused on three main categories for the comparison: orthography, lexis, and grammar. The results showed similar results for both MT systems in orthography and grammar (approximately 92%). The difference was found in the case of lexis, where Google achieved better results than Bing. The limitation of the experiment was its small dataset.

Cornet et al. [26] focused on a comparison of three MT systems: Google Translate, Matecat, and Thot. The article aimed to support the creation of Dutch interface terminologies for the usage of SNOMED CT using MT. SNOMED CT contains more than 55,000 procedures with English descriptions for which human translation would require a lot of time and resources. The MT system Matecat [27] is specific with its translation memory and offers the option of selecting the field of translation. Thot [28] is an MT system based on phrases; it was trained for the experiment using Dutch and English noun phrases. The outputs of the examined MT systems were evaluated by two reviewers. The translations were considered acceptable by the reviewers when they covered the meanings of the English terms. The results of the experiment showed that the quality of all three MT systems was not good enough for use in clinical practice, and none of them translated the terms according to the translation rules of SNOMED CT. The worst results were achieved for Thot, and the results for Google Translate and Matecat were similar.

Berrichi and Mazroui [29] focused on the issue of out-of-vocabulary words. The authors integrated morphosyntactic features into NMT models and dealt with long sentences that create issues for NMT systems in the English–Arabic language pair. The experiment was realized under low- and high-resource conditions; novel segmentation techniques were

proposed to overcome the issues with long sentences. Under low-resource conditions, the BLEU scores decreased according to the sentence size. This indicated the need to shorten the sentences before translation.

Jassem and Dwojak [30] evaluated SMT and NMT models trained on a domain-specific English–Polish corpus of medium size. The authors used the corpora from the OPUS website (open parallel corpus). In the case of phrase-based SMT, the Moses toolkit was used for translation, and the NMT model was trained with the MarianNMT toolkit [20]. The results of the automatic comparison (BLEU) showed similar performance quality for both systems, and therefore, the authors decided to compare the performance manually. The evaluation set consisted of 2000 random pairs of translated sentences. The evaluators had to decide on the winning MT system output or choose a tie for each sentence pair. The results showed that the human evaluators favored the NMT approach over the SMT approach, as it had a better fluency of its output.

The above-mentioned authors dealt with the evaluation of NMT and SMT systems on various language pairs. Machine translation into Slovak is less explored, and this was the motivation for the following experiment.

3. Materials and Methods

The aim of this study is to evaluate SMT (Google Translate and mt@ec, the European Commission's MT tool) and NMT systems (Google Translate and eTranslation, formerly mt@ec) for the translation direction of English to Slovak. Slovak, one of the official EU languages, is a synthetic language containing an inflected morphology and a loose word order [31]. The linguistic evaluation was performed using state-of-the-art automatic evaluation metrics (BLEU_n).

We compared the differences between the NMT and SMT on a specific dataset. Two different MT systems were used for this purpose. The reference translation was obtained by two professional translators using our online system, OSTPERE (Online System for Translation, Post-Editing, Revision, and Evaluation), which was originally created for post-editing machine-translated documents [32–34]. The online system was built on a PHP platform and uses a MySQL database to store the text documents and their translations.

The dataset consisted of 160 journalistic texts written in English. We translated all of the examined texts using both Google Translate (SMT and NMT) and the European Commission's MT tool (mt@ec and eTranslation). The composition of the created dataset is depicted in Table 1, and the dataset is available (Supplementary Data S1) [35].

Table 1. Lexico-grammatical dataset composition.

Feature Type	GT_SMT	GT_NMT	mt@ec_SMT	E-translation_NMT	Reference
Average sentence length (words)	21.74	19.54	18.06	20.15	18.66
Average word length (characters)	5.32	5.43	5.6	5.4	5.53
Frequency of long sentences (w $\geq$ 10)	84.83%	83.97%	73.42%	83.44%	81.87%
Frequency of short sentences (w < 10)	15.17%	16.03%	26.58%	16.56%	18.13%
Frequency of nouns	29.94%	31.00%	31.34%	29.40%	29.67%
Frequency of adjectives	9.23%	10.06%	10.25%	9.68%	10.25%
Frequency of adverbs	3.77%	3.30%	3.76%	3.56%	3.51%
Frequency of verbs	16.16%	15.02%	14.67%	16.00%	15.91%
Frequency of pronominals	7.32%	7.64%	6.88%	7.67%	8.67%
Frequency of participles	2.42%	1.81%	2.95%	2.40%	1.65%
Frequency of morphemes	1.65%	2.36%	1.61%	2.34%	3.07%
Frequency of abbreviation	0.52%	0.36%	0.88%	0.49%	0.25%
Frequency of numbers	1.74%	1.97%	1.97%	1.81%	1.33%
Frequency of undefinable POSs	0.42%	0.36%	0.39%	0.33%	0.32%
Frequency of particules	1.87%	2.20%	2.25%	2.01%	2.37%
Frequency of foreign words	3.19%	2.56%	2.21%	2.44%	1.83%
Frequency of interjections	0.00%	0.00%	0.04%	0.03%	0.00%
Frequency of numerals	3.10%	2.52%	2.74%	2.70%	2.78%
Frequency of prepositions and conjunctions	18.68%	18.83%	18.08%	19.13%	18.38%

As the focus was on the evaluation of different MT systems and architectures, we had to create an approach to the experiment. We proceeded based on a methodology consisting of the following steps:

1. Obtaining the unstructured text data (source text) of a journalistic style in English.
2. Text preparation—removing the document formatting.
3. Machine translation using various systems:
   a. Google Translate—statistical machine translation,
   b. Google Translate—neural machine translation,
   c. mt@ec—statistical machine translation,
   d. eTranslation—neural machine translation.
4. Human translation of the documents using the online system OSTPERE.
5. Text alignment—the segments of source texts were aligned with the generated MT system output, where each source text segment had its corresponding MT outputs and human translation output; this was done using the HunAlign tool [36].
6. Evaluation of the texts using automatic metrics of accuracy from the Python Natural Language Toolkit library, or NLTK, which provides an implementation of the BLEU score using the sentence_bleu function.
   In our research, we applied automatic metrics of accuracy [14]. The metrics of accuracy (e.g., precision, recall, BLEU) are based on the closeness of the MT output (h) with the reference (r) in terms of n-grams; their lexical overlap is calculated in (A) the number of common words ($h \cap r$), (B) the length (number of words) of the MT output, and (C) the length (number of words) of the reference. The higher the values of these metrics, the better the translation quality.
   BLEU (Bilingual Evaluation Understudy) [14], which is used in our research, is a geometric mean of n-gram precisions ($p = A/B$), and the second part is a brevity penalty (BP), i.e., a length-based penalty to prevent very short sentences as compensation for inappropriate translation. BLEU represents two features of MT quality—adequacy and fluency [7].

$$BLEU\_n = exp \sum_{n=1}^{N} w_n \log p_n \times BP,$$

$$p_n = precision_n = \frac{\sum_{S \in C} \sum_{n-gram \in S} count_{matched}(n-gram)}{\sum_{S \in C} \sum_{n-gram \in S} count(n-gram)},$$

$$\text{and } = \begin{cases} 1, & if\ h > r \\ e^{1-\frac{r}{h}}, & if\ h \leq r \end{cases},$$

   where S indicates hypothesis (h) and reference (r) in the complete corpus C.
7. Comparison of the translation quality based on the system (GT, EC) and translation technology (SMT, NMT).
   We verified the quality of the machine translations using the BLEU_n (n = 1, 2, 3, and 4) automatic evaluation metrics. We tested the differences in MT quality—represented by the score of the BLEU_n automatic metrics—between the translations generated from Google Translate (GT_SMT and GT_NMT) and the European Commission's MT tool (EC_mt@ec and EC_e-translation). This resulted in the following global null hypotheses:
   The quality of machine translation (BLEU_n, n = 1, 2, 3, and 4) does not depend on the MT system (GT or EC) or the translation technology (SMT or NMT).
   To test for differences between dependent samples (BLEU_n: EC_SMT, GT_SMT, EC_NMT, and GT_NMT), we used adjusted univariate tests for repeated measures due to the failure of the sphericity assumption (Mauchley sphericity test—BLEU_1: $W = 0.831$, *Chi-Square* = 29.149, $df = 5$, $p < 0.001$; BLEU_2: $W = 0.725$, *Chi-Square* = 50.651, $df = 5$, $p < 0.001$; BLEU_3: $W = 0.804$, *Chi-Square* = 34.495, $df = 5$, $p < 0.001$; BLEU_4: $W = 0.618$, *Chi-Square* = 76.023, $df = 5$, $p < 0.001$). For all BLEU_n automatic metrics, the test is significant ($p < 0.001$), i.e., the assumption is violated. If the assumption of covariance matrix sphericity is not met, the type I error rate increases. We adjusted

the degrees of freedom using the Greenhouse–Geisser adjustment ($df1 = (T-1)$, $df2 = (T-1)(D-1)$) for the F-test that was used, thus achieving the declared level of significance.

$$adj.df1 = \hat{\varepsilon}(T-1),$$

$$adj.df2 = \hat{\varepsilon}(T-1)(D-1),$$

where T is the number of dependent samples (BLEU_n scores of the examined translations) and D is the number of cases (documents). After rejecting the global null hypotheses for individual BLEU_n, the Bonferroni adjustment was used as a post hoc test [37,38].

8. Identification of the texts with the greatest distance between SMT and NMT [39]. We used residuals to compare scores of the BLEU_n automatic metrics of MT_SMT with MT_NMT at the document level. In our case, the residual analysis was defined as follows: $(residual\ value)_i = (BLEU\_n\ score\ of\ NMT\ text)_i - (BLEU\_n\ score\ of\ SMT\ text)_i\ i = 1, 2, \ldots, D$,
where D is the number of examined texts in the dataset, *NMT* is a neural machine translation, and *SMT* is a statistical machine translation.

4. Results

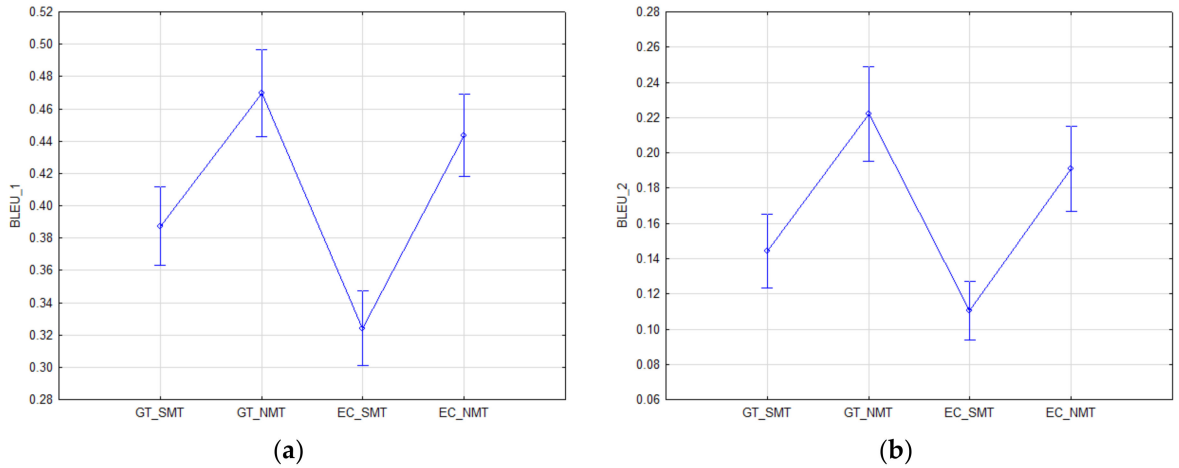
4.1. Comparison of MT Quality

Based on the results of the adjusted univariate tests for repeated measures (Greenhouse–Geisser adjustment) among GT_SMT, GT_NMT, mt@ec_SMT, and eTranslation_NMT, there are significant differences in the MT quality in terms of the scores for BLEU_n (n = 1, 2, 3 and 4: *G-G Epsilon* < 0.898, $p < 0.001$).

From multiple comparisons (Bonferroni adjustment) in the case of BLEU_1 (Table 2a), there is a significant difference between NMT (GT_NMT or eTranslation_NMT) and GT_SMT, as well as mt@ec_SMT ($p < 0.001$), in favor of NMT (Figure 1a). Similar results (Table 3a, Figure 2a) were achieved in the case of the BLEU_3 measure ($p < 0.01$).

Table 2. Multiple comparisons: (**a**) BLEU_1 and (**b**) BLEU_2.

(a)					(b)				
BLEU_1	**GT_SMT**	**GT_NMT**	**EC_SMT**	**EC_NMT**	**BLEU_2**	**GT_SMT**	**GT_NMT**	**EC_SMT**	**EC_NMT**
GT_SMT		0.00000	0.00000	0.00000	GT_SMT		0.00000	0.01047	0.00010
GT_NMT	0.00000		0.00000	0.10998	GT_NMT	0.00000		0.00000	0.02446
EC_SMT	0.00000	0.00000		0.00000	EC_SMT	0.01047	0.00000		0.00000
EC_NMT	0.00000	0.10998	0.00000		EC_NMT	0.00010	0.02446	0.00000	

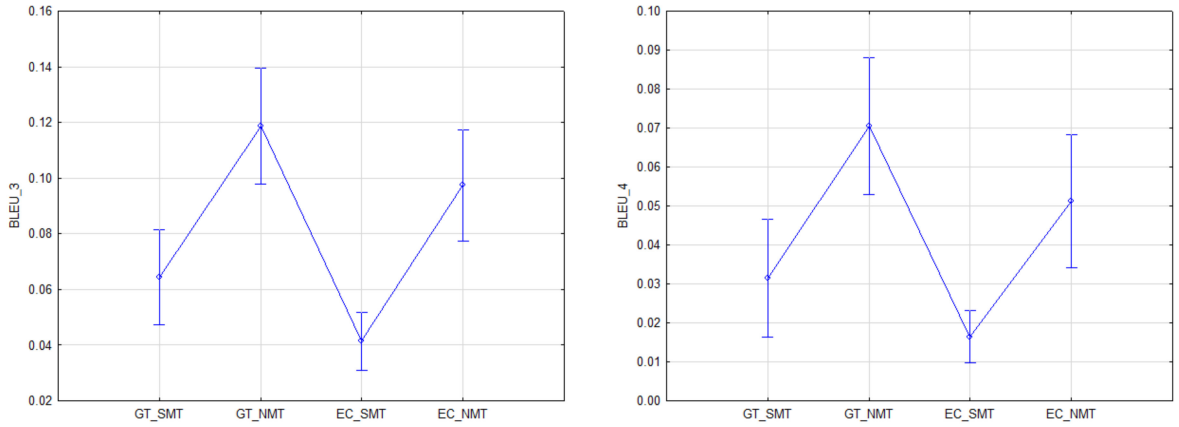


(**a**) (**b**)

Figure 1. Mean estimations of the points and intervals for the automatic metrics: (**a**) BLEU_1 and (**b**) BLEU_2.

Table 3. Multiple comparisons—Bonferroni adjustment: (**a**) BLEU_3 and (**b**) BLEU_4.

(a)					(b)				
BLEU_3	**GT_SMT**	**GT_NMT**	**EC_SMT**	**EC_NMT**	**BLEU_4**	**GT_SMT**	**GT_NMT**	**EC_SMT**	**EC_NMT**
GT_SMT		0.00000	0.07270	0.00233	GT_SMT		0.00004	0.46602	0.12269
GT_NMT	0.00000		0.00000	0.12649	GT_NMT	0.00004		0.00000	0.14572
EC_SMT	0.07270	0.00000		0.00000	EC_SMT	0.46602	0.00000		0.00030
EC_NMT	0.00233	0.12649	0.00000		EC_NMT	0.12269	0.14572	0.00030	



(**a**) (**b**)

Figure 2. Mean estimations of the points and intervals for the automatic metrics: (**a**) BLEU_3 and (**b**) BLEU_4.

The BLEU_2 measure (Table 2b) showed not only significant differences between statistical machine translation and neural machine translation ($p < 0.001$) in favor of NMT (Figure 1b), but also a significant difference between neural machine translations themselves, i.e., between GT_NMT and eTranslation_NMT ($p < 0.05$) in favor of GT_NMT.

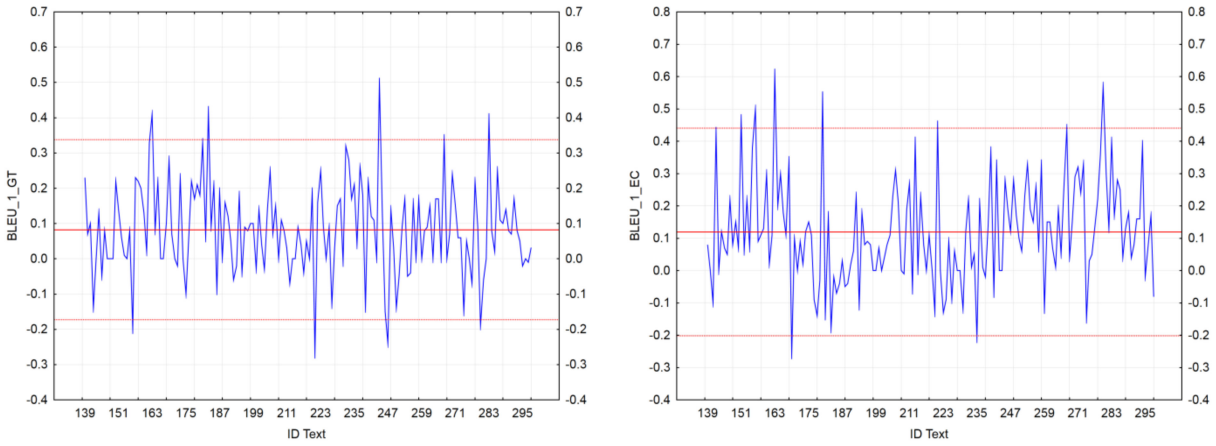
Concerning the BLEU_4 measure (Table 3b), there is a significant difference between GT_NMT and the statistical machine translations (GT_SMT or mt@ec_SMT) ($p < 0.001$) in favor GT_NMT (Figure 2b), but there is not a difference between neural machine translation (eTranslation_NMT) and statistical machine translation (GT_SMT) ($p > 0.05$).

To sum up, the assumption concerning the better quality of NMT regardless of the system used (GT_NMT or eTranslation_NMT) has been confirmed. There were statistically significant differences between SMT and NMT in favor of NMT based on all BLEU_n metrics.

Subsequently, the results were also verified using error rate metrics (CDER, PER, TER, and WER). Based on adjusted univariate tests for repeated measures and multiple comparisons, statistically significant differences between the NMT and SMT translation technologies were proved in favor of NMT (CDER, PER, TER, and WER: *G-G Epsilon* < 0.944, $p < 0.001$). The NMTs achieved statistically significantly lower error rates than the SMTs, regardless of the MT system. The results matched, and we can consider them robust.

4.2. MT Text Identification

Based on the previous results (Section 4.1), we further examined the MT texts in which the highest differences in the BLEU_n scores given to the MT systems (GT_SMT, GT_NMT, mt@ec_SMT, and eTranslation_NMT) used for translation from English into Slovak were found. To identify these texts, residuals were used [39,40]. To identify extreme values (Figure 3a,b, Figure 4a,b, Figure 5a,b, and Figure 6a,b) we used a rule of $\pm 2\sigma$, i.e., values outside the interval were considered as extremes.

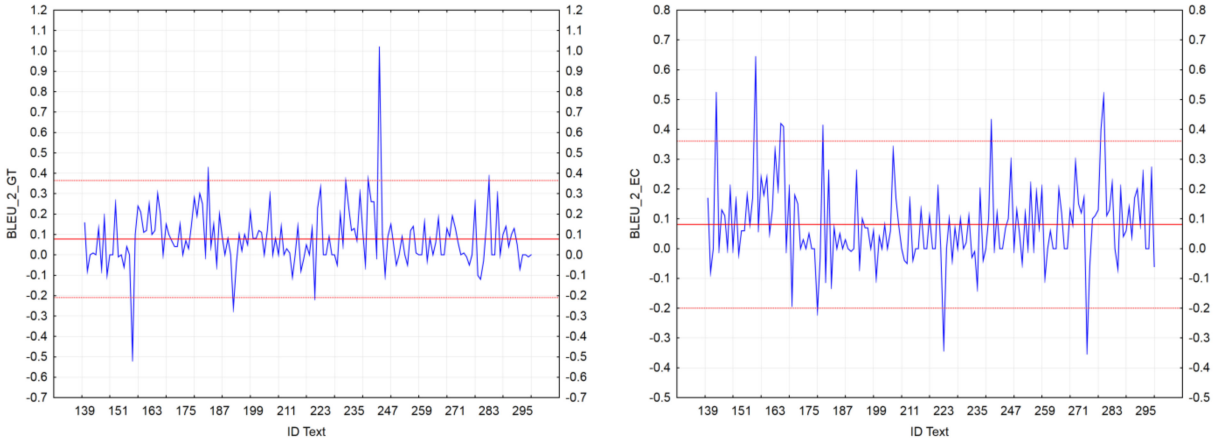


(**a**) (**b**)

Figure 3. Visualization of BLEU_1 residuals for the machine translation (MT) outputs: (**a**) Google Translate; (**b**) the European Commission's (EC's) tool.

In the case of BLEU_1, we identified five texts (ID_163, ID_183, ID_244, ID_267, and ID_283) that showed a significant accuracy of GT_NMT compared to GT_SMT; vice-versa, only four texts (ID_156, ID_221, ID_247, and ID_280) showed a significant accuracy of GT_SMT compared to GT_NMT (Figure 3a). In the case of the EC's tool, we achieved different results. Based on BLEU_1, we identified six texts (ID_151, ID_156, ID_163, ID_180, ID_221, and ID_280) that showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT; vice-versa, only two texts (ID_169 and ID_235) showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT (Figure 3b).

In the case of BLEU_2, we identified four texts (ID_183, ID_240, ID_244, and ID_283) that showed a significant accuracy of GT_NMT compared to GT_SMT; vice-versa, only two texts (ID_156 and ID_192) showed a significant accuracy of GT_SMT compared to GT_NMT (Figure 4a). In the case of the EC's tool, we achieved the following results: We identified eight texts (ID_142, ID_156, ID_165, ID_166, ID_180, ID_240, ID_279, and ID_280) that showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT; vice-versa, only three texts (ID_178, ID_223, and ID_274) showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT (Figure 4b).

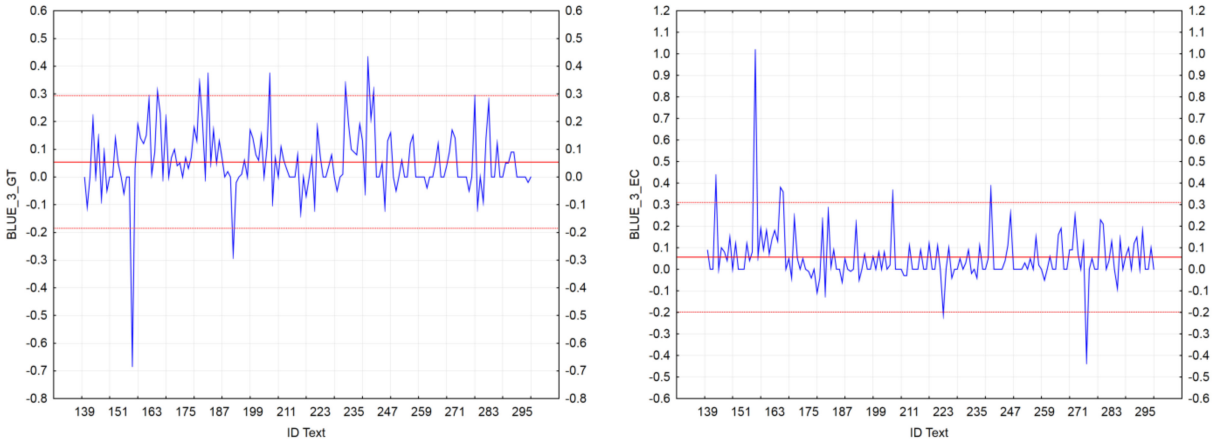


(**a**) (**b**)

Figure 4. Visualization of BLEU_2 residuals for the MT outputs: (**a**) Google Translate; (**b**) the EC's tool.

Based on BLEU_3, we identified seven texts (ID_165, ID_180, ID_183, ID_205, ID_232, ID_240, and ID_242) that showed a significant accuracy of GT_NMT compared to GT_SMT; vice-versa, only two texts (ID_156 and ID_192) showed a significant accuracy of GT_SMT compared to GT_NMT (Figure 5a). In the case of the EC's tool, we identified six texts

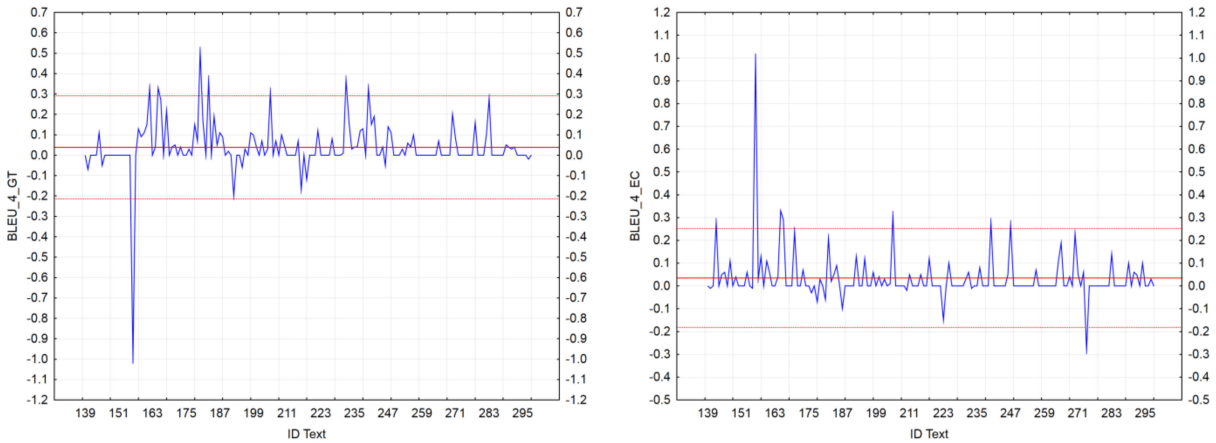
(ID_142, ID_156, ID_165, ID_166, ID_205, and ID_240) that showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT; vice-versa, only two texts (ID_223 and ID_274) showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT (Figure 5b).



(**a**) (**b**)

Figure 5. Visualization of BLEU_3 residuals for the MT outputs: (**a**) Google Translate; (**b**) the EC's tool.

In the case of BLEU_4, we identified seven texts (ID_162, ID_165, ID_180, ID_183, ID_205, ID_232, and ID_240) that showed a significant accuracy of GT_NMT compared to GT_SMT; vice-versa, only one text (ID_156) showed a significant accuracy of GT_SMT compared to GT_NMT (Figure 6a). In the case of the EC's tool, we identified seven texts (ID_142, ID_156, ID_165, ID_166, ID_205, ID_240, and ID_247) that showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT; vice-versa, only one text (ID_274) showed a significant accuracy of eTranslation_NMT compared to mt@ec_SMT (Figure 6b).



(**a**) (**b**)

Figure 6. Visualization of BLEU_4 residuals for the MT outputs: (**a**) Google Translate; (**b**) the EC's tool.

To sum up, the results of the residual analyses of the BLEU_n metrics of the neural and statistical MTs confirmed the previous results (Section 4.1). The residuals' mean was positive for all BLEU_n metrics, which proves a higher degree of accuracy in favor of NMT.

The identified texts were submitted to linguistic analysis and are described within the discussion.

5. Discussion

As the results showed, in the case of Google Translate and based on the BLEU_1 metric, neural machine translation proved to be significantly better for six texts compared to its predecessor, statistical machine translation. In the more detailed linguistic analysis, it was mainly about increasing the accuracy of the transfer of the meaning of the source text and reducing the literal translations and untranslated words. Examples of sentences are:

(1)

Source: He is believed to have stayed for a time in Belgium but has never been found.

GT_SMT: On je veril k zdržiavala na nejaký čas v Belgicku, ale nebolo nikdy nájdené.

GT_NMT: Predpokladá sa, že nejaký čas zostal v Belgicku, ale nikdy ho nenašli.

Reference: Predpokladá sa, že istý čas v Belgicku zostal, no polícia ho nikdy nevypátrala.

In GT_SMT, the English structure "he is believed to have stayed" was translated literally. The passive form "is believed" was not recognized; the past participle form was translated as the past-tense form, and the verb "is" (representing the present tense) was followed by a verb in the past tense (which sounds illogical). The issue of literal machine translation is discussed in [41].

The GT_NMT was correct and fluent, although the reference explicitly suggests the noun "polícia" (the police).

(2)

Source: The flexible partner will eventually adapt their entire life around the inflexible partner's insistences.

GT_SMT: Flexibilné partner bude nakoniec prispôsobiť celý svoj život okolo insistences nepoddajná partnera.

GT_NMT: Flexibilný partner nakoniec prispôsobí celý svoj život požiadavkám nepružného partnera.

Reference: Nakoniec flexibilný partner prispôsobí celý svoj život požiadavkám neflexibilného partnera.

We can observe some mistakes that are typical for SMT outputs: There is a low agreement between the words "flexibilné" and "partner" (the correct form is "flexibiln-ý partner"). The form "flexibiln-é" refers to the neuter gender, and the noun "partner" refers to the male gender.

The same issue is detected in the form "nepoddajná partnera". GT_SMT suggests using the form "nepoddajn-á", which is in the female gender, and the noun "partner-a", which is in the male gender. In addition to the discrepancy in the agreement in the gender, we can see the discrepancy in the category of the case (the form "*nepoddajn-á*" is in the nominative case, and "partner-a" is in the genitive or accusative case). The agreement in the prediction—namely in the predicative and agreement categories—can be crucial in the comprehensibility of a text [42].

The issues of untranslated lexemes (e.g., "insistences" in this segment) are typical for SMTs; omitted lexemes are more common for NMTs [41].

The GT_NMT was adequate and fluent. In comparison with the reference, there was just a slight adjustment in the word order, which was not necessary.

However, an interesting finding was that only one text (ID_183) had a significantly more accurate translation in terms of all four metrics (BLEU_1 to BLEU_4).

Regarding the BLEU_2 metric, in only four texts, NMT performed better than SMT. An interesting finding is that there was also a text (ID_240) in which the same quality improvement was proven by BLEU_ 3 (increase in the adequacy and fluency). This text contained sentences such as:

(3)

Source: How much should we fear the rise of artificial intelligence?

GT_SMT: Koľko by sme mali báť vzostup umelej inteligencie?

GT_NMT: Nakoľko by sme sa mali báť nárastu umelej inteligencie?

Reference: Nakoľko by sme sa mali obávať vzostupu umelej inteligencie?

The GT_SMT points out the issue of the incorrect transfer of words. The errors occur due to the different typological characters of English and Slovak. This can be explained by the fact that homonymy and polysemy are typical for languages with poor derivation (e.g., English). Slovak has a richer derivation of words and a lower number of polysemous and homonymous words [43]. In such cases, MT is not able to choose a meaning that is adequate for a particular context from a large range of homonymous and polysemantic words. In the GT_SMT, the phrase "how much" can be translated as "koľko" as well as "nakoľko". Such errors can be revised by a human post-editor. The GT_NMT was adequate and fluent.

The BLEU_4 metric has a specific nature; it is a match of four consecutive tokens, regardless of whether they are words or punctuation. The following text can serve as an example:

(4)

Source: He said they would facilitate "speedier, clearer investigations and stricter prison sentences" and would help the authorities understand Isis's "underlying structures".

GT_SMT: Hovoril, že by to uľahčilo "rýchlejšie, čistejšie vyšetrovanie a prísnejšie tresty odňatia slobody" a bol by pomohla orgánom pochopiť ISIS je "základnej štruktúry".

(Back translation: He said they would facilitate "speedier, clearer investigations and stricter prison sentences" and would help the authorities understand Isis is "underlying structures".)

GT_NMT: Povedal, že by uľahčili „rýchlejšie a jasnejšie vyšetrovanie a prísnejšie tresty odňatia slobody"a pomôžu úradom pochopiť „základné štruktúry"Isis.

(Back translation: He said they would facilitate "speedier, clearer investigations and stricter prison sentences" and would help the authorities understand Isis's "underlying structures".)

Reference: Povedal, že budú presadzovať „rýchlejšie, jasnejšie vyšetrovania a prísnejšie tresty odňatia slobody", a pomôžu úradom pochopiť „základné štruktúry Isis".

In GT_SMT, we can observe a literal translation, especially in the possessive form "Isis's", which was confusingly translated as a short form of the verb "is/has".

Our research results also pointed to the opposite phenomenon, i.e., to texts in which neural machine translation simultaneously achieved a lower score for all BLEU_n metrics than its predecessor (GT_SMT)—specifically, in texts that contained sentences with extra words, for example:

(5)

Source: The search for suspects continues.

GT_SMT: Pátranie po podozrivých pokračuje.

GT_NMT: Pátranie po podozrivých osobách pokračuje.

Reference: Pátranie po podozrivých pokračuje.

The outputs of both GT_SMT and GT_NMT were adequate, correct, and fluent; GT_NMT offered the explication "osobách" (persons), which was not required.

In the case of the EC's tool, the results were similar, but not as significantly different as in the case of GT. In six texts, the neural machine translation achieved a better score for the BLEU_1 metric than its predecessor. This was mainly an increase in the quality of translation, i.e., in the adequacy of the transfer of the source text into the target language:

(6)

Source: Among their number were Belgian students, French schoolchildren and British lawyers. mt@ec_SMT: Spomedzi nich boli belgické, francúzske a britské študentov, advokátov.

(Back translation: Among their number were Belgian, French and British students, lawyers.)

eTranslation_NMT: Medzi ich počet boli belgickí študenti, francúzski žiaci a britskí právnici.

(Back translation: Among their number were Belgian students, French pupils and British lawyers.)

Reference: Nachádzajú sa medzi nimi belgickí študenti, francúzski školáci a britskí právnici.

In GT_SMT, we can observe the issues of agreement in the case and gender ("belgické, francúzske a britské študentov") and the issue of the adequate transfer of the word "among" (in this context, "medzi nimi" is a better option than "spomedzi nich"). Moreover, some lexemes in the syntagms "Belgian students, French schoolchildren and British lawyers" were omitted; the syntagms were broken, and they were translated incorrectly.

Similar results were shown for the BLEU_2 metric. The neural machine translation achieved a better translation quality due to the reference. Its translation was more accurate in both meaning and fluency; for example:

(7)

Source: Michel called on residents to "stay calm and cool-headed" as the investigation continued into Tuesday's police raid.

mt@ec_SMT: Michel vyzval, aby zachovali pokoj a „cool-headed"ako prešetrovanie pokračovalo utorkového policajným zásahom.

eTranslation_NMT: Michel vyzval obyvateľov, aby „zostali pokojní a chladnohlaví", keď vyšetrovanie pokračovalo v utorkovom policajnom nájazde.

Reference: Po utorkovom policajnom zásahu Michel vyzval obyvateľov, aby „zostali pokojní a nepodliehali panike."

GT_SMT omitted two lexemes ("residents", "cool-headed") and numbered several errors in the agreement in the number and case ("utorkového policajným zásahom"). With these errors, the GT_SMT was inadequate and incomprehensible.

In the GT_NMT, we can see the improvement in MT—in comparison to GT_SMT—in punctuation. Slovak uses a different method of quotation [44].

The BLEU_3 metric showed very similar results to those for the BLEU_2 metric. Again, the translation was more accurate than statistical machine translation due to the reference. This was mainly in sentences such as:

(8)

Source: Three officers were injured during the operation.

mt@ec_SMT: Troch úradníkov boli zranené počas prevádzky.

(Back translation: Of three officers were injured during the operation.)

eTranslation_NMT: Počas operácie boli zranení traja dôstojníci.

(Back translation: Three officers were injured during the operation.)

Reference: Počas operácie boli zranení traja policajti.

Some sentences contained identical neural machine translations as a reference, i.e., all BLEU_n metrics showed higher scores in the neural machine translation than in the statistical machine translation; e.g.,

(9)

Source: The search for suspects continues.

mt@ec_SMT: Vyhľadávanie podozrivých pokračuje.

eTranslation_NMT: Pátranie po podozrivých pokračuje.

Reference: Pátranie po podozrivých pokračuje.

Both sentences are fluent and comprehensible; GT_SMT used a less adequate equivalent for the word "search" than GT_NMT.

Similarly to GT, in the EC's tool, we also recorded a decreasing quality of the neural machine translation compared to the statistical MT. In addition the erroneous declension, untranslated words occurred in the NMT, but such a significant difference in the quality of the translation as in the GT was not proven; e.g.,

(10)

Source: The names of several people linked to last November's Paris terror attacks appear in a cache of leaked Islamic State documents, according to reports in Germany.

mt@ec_SMT: Mená viacerých osôb spojených s minulý rok v novembri v Paríži teroristických útokov sú vo vyrovnávacej úniku Islamského štátu, podľa správ v Nemecku.

eTranslation_NMT: Mená niekoľkých ľudí spojených s Paríž teroristických útokov november minulého roka je uvedené cache z uniknutých dokumentov islamský štát, podľa správ v Nemecku.

Reference: Podľa správ z Nemecka sa mená viacerých osôb spájaných s novembrovými teroristickými útokmi v Paríži objavujú v mnohých odhalených dokumentoch Islamského štátu.

The issues in the category of agreement in gender, number, and case appeared even in the last example (10). We agree with [41] that the situation with translation worsens with multi-word sentence elements and complicated sentence structures (e.g., a group of southern Pacific islands, Europe's centuries-long history, local agricultural business leaders, a positive contribution to developing an effective post-Brexit immigration policy).

Through our analysis, we not only verified our assumption about the better quality of neural machine translation, but mainly through the residual analysis, we were also able to identify the texts with the greatest distance between NMT and SMT. This allowed us to focus more precisely—in terms of more detailed linguistic analysis—on texts with the greatest distances in translation accuracy between examined translation technologies. The present methodology is repeatable for any language pairs, translation technologies, and MT systems using different metrics and measures of MT evaluation.

6. Conclusions

Our research aimed to establish whether machine translation based on neural networks achieves a higher quality than its predecessor, statistical machine translation, in terms of translation accuracy.

We can summarize the results of our research into two interesting contributions. The first contribution is that neural machine translation achieves a better quality of translation of journalistic texts from English into Slovak, whether it uses a system trained on general texts (Google Translate) or a specific one that is trained on a specific domain (the EC's tool). This study—a linguistic evaluation of MT output—is unique in the examined direction of translation. Although there are studies in which the Slovak language was used, these studies were focused on training MT systems using several languages, e.g., for training massively multilingual NMT models or adapting neural machine translation systems to new, low-resource languages [45,46]. Machine translation into Slovak has seldom been explored in contrast to Czech, which is genetically very close to Slovak, and for which several studies focusing on evaluating the quality of machine translation exist [47,48].

The second contribution is that, for machine translation of journalistic texts (mostly of the news genre) from English into Slovak, it is better to use a general MT system (GT) than a specialized one, even though is the texts are of the news genre. Of all the MT systems examined, GT_NMT achieved the highest translation quality due to its accuracy, as measured using the BLEU_n metrics and the reference. There was no significant difference in translation quality between the GT_SMT statistical machine translation and the eTranslation_NMT neural machine translation due to BLEU_4.

Our results correspond to the findings of [1] which show that it is possible to obtain 70–80% accuracy in machine translation using artificial intelligence (multilingual NMT models); however, the rest is still a task for human translation.

The future direction of our research will consist of expanding the examined texts and using other automatic metrics for evaluating the quality of translation, such as HTER (Human-targeted Translation Error Rate), which is based on edit distance.

Supplementary Materials: The following are available online at http://dx.doi.org/10.17632/r6f44z6ycf.1 (accessed on 28.1.2021): Dataset S1: data.txt.

Author Contributions: Conceptualization, L.B. and M.M.; methodology, Ľ.B.; validation, D.M. and M.M.; formal analysis, D.M. and M.M.; investigation, L.B.; resources, L.B. and Ľ.B.; data curation, Ľ.B.; writing—original draft preparation, L.B. and Ľ.B.; writing—review and editing, D.M. and M.M.;

visualization, M.M.; supervision, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and the Slovak Academy of Sciences (SAS) under contract No. VEGA-1/0809/18, as well as by the scientific research project of the Czech Sciences Foundation Grant No.: 19-15498S and by the Slovak Research and Development Agency under contract no. APVV-18-0473.

Data Availability Statement: The data presented in this study are openly available in [Mendeley Data] at [doi:10.17632/r6f44z6ycf.1], reference number [35].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Farhan, W.; Talafha, B.; Abuammar, A.; Jaikat, R.; Al-Ayyoub, M.; Tarakji, A.B.; Toma, A. Unsupervised dialectal neural machine translation. *Inf. Process. Manag.* **2020**, *57*, 102181. [CrossRef]
2. Chomsky, N. Three Factors in Language Design. *Linguist. Inq.* **2005**, *36*, 1–22. [CrossRef]
3. Christensen, C.H. Arguments for and against the Idea of Universal Grammar. *Leviathan Interdiscip. J. Engl.* **2019**, *4*, 12–28. [CrossRef]
4. Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. Approaches to Human and Machine Translation Quality Assessment. In *Translation Quality Assessment. Machine Translation: Technologies and Applications*; Springer: Cham, Switzerland, 2018; Volume 1.
5. Popović, M. Error Classification and Analysis for Machine Translation Quality Assessment. In *Machine Translation: Technologies and Applications*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S., Eds.; Springer: Cham, Switzerland, 2018; Volume 1.
6. Dowling, M.; Moorkens, J.; Way, A.; Castilho, S.; Lynn, T. A human evaluation of English-Irish statistical and neural machine translation. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 3–5 November 2020; European Association for Machine Translation: Lisboa, Portugal, 2020; pp. 431–440.
7. 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.* **2018**, *34*, 3225–3233. [CrossRef]
8. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
9. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, 6–11 August 2017; Volume 3.
10. Zhang, J.J.; Zong, C.Q. Neural machine translation: Challenges, progress and future. *Sci. China Technol. Sci.* **2020**, *63*, 2028–2050. [CrossRef]
11. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, 3104–3112.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates: Long Beach, CA, USA, 2017; Volume 2017-Decem.
13. Biesialska, M.; Guardia, L.; Costa-jussa, M.R. *The TALP-UPC System for the WMT Similar Language Task: Statistical vs Neural Machine Translation*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 185–191.
14. 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, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
15. Webster, R.; Fonteyne, M.; Tezcan, A.; Macken, L.; Daems, J. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. *Informatics* **2020**, *7*, 21. [CrossRef]
16. Van Brussel, L.; Tezcan, A.; Macken, L. A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch. In Proceedings of the LREC 2018-11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
17. Tezcan, A.; Daems, J.; Macken, L. When a 'sport' is a person and other issues for NMT of novels. In *Qualities of Literary Machine Translation*; European Association for Machine Translation: Dublin, Ireland, 2019; pp. 40–49.
18. Yu, H.; Luo, H.; Yi, Y.; Cheng, F. A2R2: Robust Unsupervised Neural Machine Translation With Adversarial Attack and Regularization on Representations. *IEEE Access* **2021**, *9*, 19990–19998. [CrossRef]
19. Haque, R.; Hasanuzzaman, M.; Way, A. Analysing terminology translation errors in statistical and neural machine translation. *Mach. Transl.* **2020**, *34*, 149–195. [CrossRef]
20. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Heafield, H.H.K.; Neckermann, T.; Seide, F.; Germann, U.; Aji, A.F.; Bogoychev, N.; Martins, A.F.T.; et al. Marian: Fast neural machine translation in c++. In Proceedings of the ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations, Toronto, ON, Canada, 31 July 2018.

21. Denkowski, M.; Lavie, A. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Sixth Workshop on Statistical Machine Translation*; Association for Computational Linguistics: Edinburgh, Scotland, 2011; pp. 85–91.
22. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. *A Study of Translation Edit Rate with Targeted Human Annotation*; Association for Machine Translation in the Americas: East Stroudsburg, PA, USA, 2006; pp. 223–231.
23. Dashtipour, K.; Gogate, M.; Li, J.; Jiang, F.; Kong, B.; Hussain, A. A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. *Neurocomputing* **2020**, *380*, 1–10. [CrossRef]
24. Almahasees, Z.M. Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. *Int. J. Lang. Lit. Linguist.* **2017**, *3*, 1–4. [CrossRef]
25. Almahasees, Z.M. Assessment of Google and Microsoft Bing Translation of Journalistic Texts. *Int. J. Lang. Lit. Linguist.* **2018**, *4*, 231–235. [CrossRef]
26. Cornet, R.; Hill, C.; De Keizer, N. Comparison of three english-to-Dutch machine translations of SNOMED CT procedures. In *Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2017; Volume 245, pp. 848–852.
27. Federico, M.; Bertoldi, N.; Cettolo, M.; Negri, M.; Turchi, M.; Trombetti, M.; Cattelan, A.; Farina, A.; Lupinetti, D.; Martines, A.; et al. The MateCat Tool. In Proceedings of the COLING 2014, 25th International Conference on Computational Linguistics: System Demonstrations, Dublin, Ireland, 23–29 August 2014; pp. 129–132.
28. Ortiz-Martínez, D.; Casacuberta, F. The New Thot Toolkit for Fully-Automatic and Interactive Statistical Machine Translation. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; Association for Computational Linguistics: Gothenburg, Sweden, 2014; pp. 45–48.
29. Berrichi, S.; Mazroui, A. Addressing Limited Vocabulary and Long Sentences Constraints in English–Arabic Neural Machine Translation. *Arab. J. Sci. Eng.* **2021**, *1744*, 1–4. [CrossRef]
30. Jassem, K.; Dwojak, T. Statistical versus neural machine translation - a case study for a medium size domain-specific bilingual corpus. *Pozn. Stud. Contemp. Linguist.* **2019**, *55*, 491–515. [CrossRef]
31. Kosta, P. Targets, Theory and Methods of Slavic Generative Syntax: Minimalism, Negation and Clitics. Slavic Languages. Slavische Sprachen. An International Handbook of their Structure. In *Slavic Languages. Slavische Sprachen. An International Handbook of their Structure, their History and their Investigation. Ein internationales Handbuch ihrer Struktur, ihrer Geschichte und ihrer Erforschung*; Kempgen, S., Kosta, P., Berger, T., Gutschmidt, K., Eds.; Mouton. de Gruyter: Berlin, Germany; New York, NY, USA, 2009; pp. 282–316. ISBN 978-3-11-021447-5.
32. Munková, D.; Munk, M.; Benko, Ľ.; Absolon, J. From Old Fashioned "One Size Fits All" to Tailor Made Online Training. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 916, pp. 365–376.
33. Munkova, D.; Kapusta, J.; Drlik, M. System for Post-Editing and Automatic Error Classification of Machine Translation. In Proceedings of the DIVAI 2016: 11th International Scientific Conference On Distance Learning in Applied Informatics; Turcani, M., Balogh, Z., Munk, M., Benko, L., Eds.; Wolters Kluwer: Sturovo, Slovakia, 2016; pp. 571–579.
34. Benko, Ľ.; Munková, D. Application of POS Tagging in Machine Translation Evaluation. In Proceedings of the DIVAI 2016: 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, Slovakia, 2–4 May 2016; Wolters Kluwer: Sturovo, Slovakia, 2016; pp. 471–489, ISSN 2464-7489.
35. Benkova, L.; Munkova, D.; Benko, L.; Munk, M. Dataset of evaluation metrics for journalistic texts EN/SK. *Mendeley Data* **2021**, *V1*. [CrossRef]
36. Varga, D.; Németh, L.; Halácsy, P.; Kornai, A.; Trón, V.; Nagy, V. Parallel corpora for medium density languages. *Proc. RANLP* **2005**, *4*, 590–596.
37. Lee, S.; Lee, D.K. What is the proper way to apply the multiple comparison test? *Korean J. Anesthesiol.* **2018**, *71*, 353–360. [CrossRef] [PubMed]
38. Genç, S.; Soysal, M.İ. Parametrik Ve Parametrik Olmayan Çoklu Karşilaştirma Testleri. *Black Sea J. Eng. Sci.* **2018**, *1*, 18–27.
39. Munk, M.; Munkova, D. Detecting errors in machine translation using residuals and metrics of automatic evaluation. *J. Intell. Fuzzy Syst.* **2018**, *34*, 3211–3223. [CrossRef]
40. Munkova, D.; Munk, M. Automatic Evaluation of Machine Translation Through the Residual Analysis. In *Advanced Intelligent Computing Theories and Applications*; Huang, D.S., Han, K., Eds.; icic 2015; pt iii.; Springer: Fuzhou, China, 2015; Volume 9227, pp. 481–490.
41. Welnitzova, K. Post-Editing of Publicistic Texts in The Context of Thinking and Editing Time. In Proceedings of the 7th SWS International Scientific Conference on Arts and Humanities-ISCAH 2020, Sofia, Bulgaria, 25–27 August 2020; STEF92Technology: Sofia, Bulgaria, 2020.
42. Welnitzová, K. Interpretačná analýza chýb strojového prekladu publicistického štýlu z anglického jazyka do slovenského jazyka. In *Myliť sa je ľudské (ale aj strojové): Analýza chýb strojového prekladu do slovenčiny*; UKF: Nitra, Slovakia, 2017; pp. 89–116. ISBN 978-80-558-1255-7.
43. Welnitzova, K.; Jakubickova, B. Enhancing cultural competence in interpreting-cultural differences between the UK and Slovakia. In Proceedings of the 7th SWS International Scientific Conference on Arts And Humanities-ISCAH 2020, Sofia, Bulgaria, 25–27 August 2020; STEF92Technology: Sofia, Bulgaria, 2020.
44. Welnitzová, K. *Neverbálna komunikácia vo svetle konzekutívneho tlmočenia*; UKF: Nitra, Slovakia, 2012; ISBN 978-80-558-0077-6.

45. Neubig, G.; Hu, J. *Rapid Adaptation of Neural Machine Translation to New Languages*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
46. Aharoni, R.; Johnson, M.; Firat, O. *Massively Multilingual Neural Machine Translation*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
47. Vojtěchová, T.; Novák, M.; Klouček, M.; Bojar, O. SAO WMT19 Test Suite: Machine Translation of Audit Reports. In Proceedings of the Fourth Conference on Machine Translation-Proceedings of the Conference, Florence, Italy, 1–2 August 2019; pp. 680–692.
48. Barrault, L.; Bojar, O.; Costa-jussà, M.R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; et al. *Findings of the 2019 Conference on Machine Translation (WMT19)*; Association for Computational Linguistics (ACL): Florence, Italy, 2019; Volume 2, pp. 1–61.