Clash between Segment-level MT Error Analysis and Selected Lexical Similarity Metrics

Marija Brkic Bakaric[1], Kristina Tonkovic[2], Lucia Nacinovic Prskalo[3]
Department of Informatics
University of Rijeka
Rijeka, Croatia

***Abstract*—The aim of this paper is to evaluate the quality of popular machine translation engines on three texts of different genre in a scenario in which both source and target languages are morphologically rich. Translations are obtained from Google Translate and Microsoft Bing engines and German-Croatian is selected as the language pair. The analysis entails both human and automatic evaluation. The process of error analysis, which is time-consuming and often tiresome, is conducted in the user-friendly Windows 10 application TREAT. Prior to annotation, training is conducted in order to familiarize the annotator with MQM, which is used in the annotation task, and the interface of TREAT. The annotation guidelines elaborated with examples are provided. The evaluation is also conducted with automatic metrics BLEU and CHRF++ in order to assess their segment-level correlation with human annotations on three different levels–accuracy, mistranslation, and the total number of errors. Our findings indicate that neither the total number of errors, nor the most prominent error category and subcategory, show consistent and statistically significant segment-level correlation with the selected automatic metrics.**

***Keywords*—Machine translation; evaluation; error analysis; BLEU; CHRF++; MQM**

I. Introduction

Machine translation (MT) is used on a daily basis by millions of people and for a range of use cases [1]. Although it will not replace humans any time soon, it can be used as a tool to enhance productivity [2]. Different types of data mean variations in structure, genre, and style, and can result in MT outputs of quite different quality. In order to properly evaluate MT, all data features important for the future use of the translation system need to be covered by the evaluation set, which is comprised of a set of sentences in the source language and their target language translations [2]. After Neural Machine Translation (NMT) took over the world scene from its predecessor phrase-based statistical MT (SMT), there have been a lot of research initiatives which focus on translation error types in an attempt to better describe differences between these two approaches. However, a standard in assessing translation quality does not exist since MT quality evaluation is subjective in its nature and quality depends on the context [3]. A review of translation quality definitions is given in [4].

Unlike the usual automatic and human evaluation metrics which provide only quantitative evaluation, error analysis enables assessing translation in qualitative terms [5]. It refers to the identification and classification of individual errors in a translated text [6]. Not only that it can reveal strengths and weaknesses of MT engines [6], but it can also show whether a system is superior over any other system regarding one aspect, all aspects, or a subset of them [7]. Multidimensional quality measure (MQM) lists quality issue types which can be used for defining specific metrics for annotation tasks and quality assessment [8]. It is used for evaluating both human and MT translations and, as such, represents a way of connecting the two. A detailed error taxonomy compliant with the hierarchical listing of issue types defined as part of the MQM which is relevant to Croatian is presented in [9]. Translation Error Annotation Tool (TREAT), which employs MQM, is described and tested in [10]. It is worth noting that human evaluation using non-directly expressed judgment-based (non-DEJ-based) metrics is more objective than that of using DEJ-based metrics, and less prone to indirect comparisons of previously assessed segments [4].

A number of metrics for automatic evaluation have been proposed up to date. These metrics are generally benchmarked against manual judgments in terms of system and segment-level correlation. This is typically done in the task of ranking various MT system translations for the same source segment [11]. New automatic MT evaluation metrics constantly emerge. Moreover, there is a metrics-shared task, which is held annually at the Workshop on MT (WMT), where new evaluation metrics are proposed. Metrics can be more or less reliable, depending on the target language, text type and genre, type of MT system, properties of human translation, and the quality aspect measured [11]. High cost and irreproducibility of human judgments can be resolved by automatic evaluation only if the latter matches human evaluation, as acknowledged in [12]. For a detailed overview of metrics and their advantages and disadvantages, we refer the reader to [4]. Metrics based on neural networks have lately shown great potential [14].

Since presenting all of the metrics and calculating their scores cannot be presented in a clear and concise way due to their vast number, only two metrics are selected from the reference-based class of metrics and chosen for the purpose of this research. These metrics always give the same score to the same text, given that all the evaluation parameters stay unaltered [4]. Both of them are based on the lexical similarity between machine translations and reference translations. While the first metric we employ is the *de facto* standard in MT community, i.e. Bilingual Evaluation Understudy (BLEU) [15], the second one – Character n-gram F-score (CHRF++),

shows promising results for morphologically rich languages [16], [17]. BLEU is a precision-based metric [15], which expresses lexical similarity on a 0-1 scale, 0 being the minimum score. While BLEU does not account for the recall directly, but through the brevity penalty, CHRF++ is a combination of precision and recall. Both of these metrics aim to achieve strong negative correlation with human error assessments, unlike error metrics which aim to achieve strong positive correlation with such human assessments. Reference translation-based metrics do require translators, and hence do not allow neither full automation of the process, nor cost and time minimization [4].

In this paper we examine correlation between error analysis results and the selected automatic metrics. The paper is organized as follows. Related work is given in Section 2, followed by the description of the research methodology in Section 3. Error analysis is presented in Section 4, results in Section 5, and discussion is provided in Section 6. Section 7 concludes the paper and gives directions for future work.

II. Related Work

The evaluation conducted in WMT 2017 examines system level correlation of metrics' scores and manual rankings and segment-level correlation with manual judgments in terms of direct assessment (DA) and UCCA-based MT evaluation (HUME) [12], where UCCA stands for universal conceptual cognitive annotation [13]. A subsequent evaluation in 2019 employs only DA [14]. The authors warn that the metrics results can be overly optimistic when the underlying set of MT systems comprises both well-performing and bad-performing systems. If the sampling of sentences does not provide sufficient number of assessments of the same segment, the evaluation tasks resort to a relative ranking re-interpretation of DA scores (DARR) [12], [14].

CHRF++ is selected for this study since it shows promising results for morphologically rich languages [16], [17]. Best CHRF correlations with human rankings are achieved for 6-grams, both on system and segment level [16]. The results in [17] show that apart from character n-grams, word 1-grams (CHRF+) and 2-grams (CHRF++) also correlate rather well with DAs. The results in [12] confirm that, on average, character-based metrics outperform other metrics. However, segment-level correlations are only around 0.4 or slightly above. BLEU is outperformed not only by character-based metrics, but also by the metric developed by the US National Institute of Standards and Technology (NIST) [18] and Translation Edit Rate (TER) [19] in the scenario where only one reference translation is used instead of the recommended four [15]. The authors in [2] see future efforts in MT evaluation directed toward character-based metrics which show the highest correlation with human judgments at both system and segment levels.

Due to inconsistencies in automatic metrics reported in [20], the study presented in this paper is conducted on three short texts. The authors in [20] base their findings on DA human evaluation of outputs of three different MT systems on three problematic domains. Human judges rate each translation on how adequately it expresses the meaning of the respective reference translation. Large differences are observed in correlations between automatic metrics and human DA across different domains. A more complex corpus with longer sentences and more complex syntactic structures turns out to exhibit higher correlations between all automatic metrics and human judgments. Reference [11] shows that performance of metrics also significantly varies across different levels of MT quality. The correlations of all evaluation metrics, including BLEU and CHRF3, are substantially lower for low-quality MT output. Not only that metrics are not able to capture nuanced quality distinctions, but they perform poorly when faced with low-quality translations. Moreover, evaluating low-quality translations is challenging even for humans. In addition, metrics prove to be more reliable when evaluating neural MT, as opposed to statistical MT systems. The difference in the evaluation accuracy for different metrics is maintained even when the gold standard scores are based on different criteria.

As expected, there is not much work on MT evaluation involving both German and Croatian. Error analysis based on Vilar's taxonomy is conducted in [21] on the Croatian translation of an essay in German and on the German translation of the same essay in Croatian. Translations are generated by Google Translate (GT). The German-to-Croatian translation direction, which proves to be more difficult for MT, is assessed by two native speakers, while the other direction is assessed by a final year graduate student of German. Incorrect word proves to be the most frequent error type in both directions.

This paper examines correlation between error analysis results and the selected automatic metrics due to several reasons. Beside the fact that DA has already been thoroughly investigated, and that it is often not feasible to obtain multiple judgments for each segment, the correlation between error counts and automatic metrics has been poorly explored. We assess the performance of the selected evaluation metrics in terms of segment-level correlation with human error analysis. We opt for the segment-level evaluation since system-level evaluation is generally an easy task for MT evaluation metrics as the majority of metrics performs extremely well at ranking systems [11].

III. Methodology

GT and Microsoft Bing engine are used for obtaining Croatian translations of German sentences, which are then annotated by a human annotator following the MQM Slavic tagset [9] and using the tool TREAT [10]. The authors in [4] call this type of human evaluation a non-DEJ-based evaluation as the judgment is not expressed directly in terms of "better than", i.e. ranking, or "good", i.e. direct assessment. Respective human translations are provided for reference.

Automatic metrics employed in the paper depend on the availability of human reference translations. Since they evaluate outputs of MT systems by comparing them to reference translations, they are also called reference translation-based metrics. BLEU is chosen, despite many of its drawbacks, as it is the *de facto* standard in MT community, and CHRF++ is chosen since it represents a very promising evaluation metric especially for morphologically rich target languages [17].

Pearson correlations between human metric segment-level scores, derived from the total number of errors and the number of errors assigned to the most frequent error category and subcategory, and automatic metric segment-level scores BLEU and CHRF++ are calculated. A stronger negative correlation indicates better performance.

The remainder of the section is divided into several subsections which give descriptions of the evaluation set, the tool used for the manual error analysis, the annotator, MQM issue types used, and the metrics BLEU and CHRF++ employed in the automatic evaluation.

A. Evaluation Set

Our evaluation set consists of 54 sentences (Table I). This is admittedly a small sample. However, this type of task can quickly become tedious so we did not want to risk inconsistent evaluation or overseeing errors. Under annotation overload, one easily becomes too tired and thus less attentive. In general, when dealing with human evaluation, there is always a necessary trade-off between the size of the sample and the integrity of the results, as acknowledged by [22].

TABLE I. EVALUATION SET DESCRIPTION

	Text		
	Recipe	***Manual***	***News***
# of sentences	18	22	14
# of words	268	280	300

The evaluation set consists of three texts in German. Three short texts are taken into consideration instead of just one longer because of the differences between automatic metrics and human DA which have been detected when dealing with different domains and text types. Texts are chosen randomly, the only requirement being that they are of approximately same size. One text is extracted from a book of recipes, one from a mobile phone manual, and one is a newspaper article. The text from the book of recipes is actually compiled of two recipes. The second text is on battery saving and battery charging. The newspaper article is about the meeting between the German chancellor, Angela Merkel, and the US president, Donald Trump.

The annotator is presented with German source texts, their respective Microsoft Bing and GT translations into Croatian, and reference translations. Since MT engines show a constant improvement over time, the translations obtained only a month later might differ greatly and might be of much better quality.

B. TREAT Application

The Universal Windows Platform (UWP) application developed and presented in [10] is chosen for the manual MQM annotation task. The user interface is shown in Fig. 1. The error analysis annotation process in TREAT is shown in Fig. 2. The input to the annotation process are three textual files – the source file, the target file, and the optional reference file.

C. Annotator

The annotator is a native speaker of Croatian with a BA degree in the German language and being very confident about her knowledge of German.

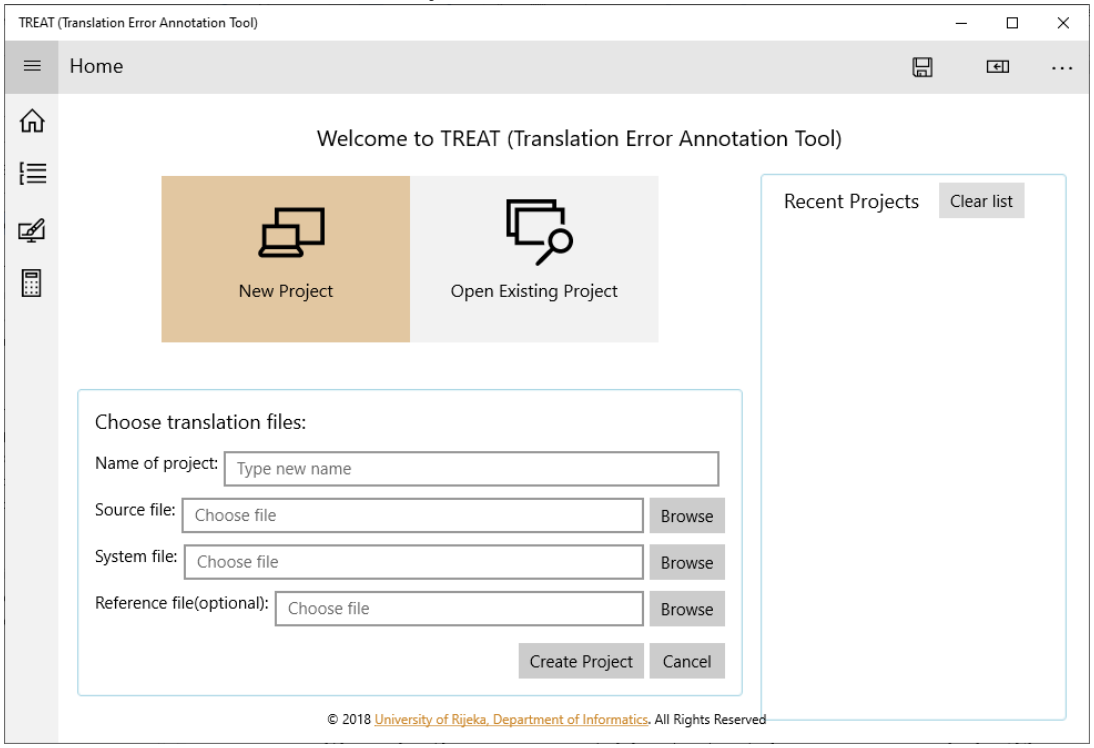

Fig. 1. TREAT user Interface.

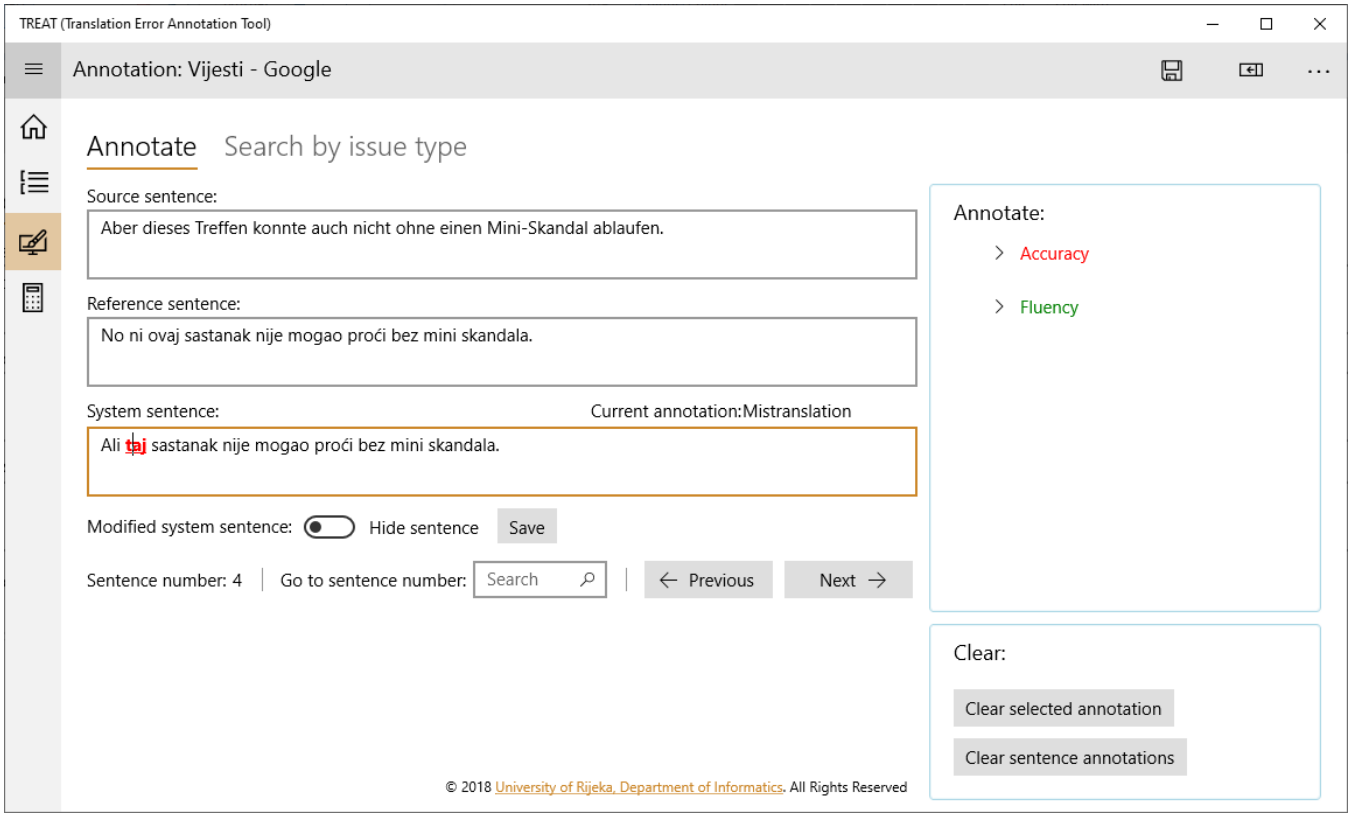

Fig. 2. Error Analysis in TREAT.

D. MQM

MQM defines over 100 issue types. The term issue is used to refer to any potential error detected in a text. At the top level there are 10 categories: accuracy, design, fluency, internationalization, locale convention, style, terminology, verity, compatibility, and other. Since it would not be viable to perform the annotation process using the full MQM tag set, it is necessary to choose a smaller subset of interest. The annotator is instructed to use the Slavic tagset [9] of the MQM core with a modification of using typography, as suggested by the core, instead of register suggested by [9]. The Slavic tagset entails higher-level categories accuracy and fluency. While accuracy is included in its original form, the authors suggest three subcategories of the word form issue–part of speech, agreement, and tense/aspect/mood, and three subcategories of function words–extraneous, missing, and incorrect. Typography refers to the issues related to the mechanical presentation of a text (e.g. punctuation is used incorrectly or a text has an extraneous hard return in the middle of a paragraph). This category should be used for any typographical errors other than spelling. If the exact nature of the error cannot be determined and a major break down in fluency occurs, unintelligible mark-up should be used.

Prior to annotation, the annotator is familiarized with TREAT and the official MQM annotation guidelines, which offer detailed instructions for annotation within MQM[1]. The annotator is instructed to avoid guessing by choosing rather a higher-level issue and to use a minimalistic mark-up. Training, evaluation guidelines elaborated with examples, and familiarity with the field to which the text belongs are considered important for evaluation [4].

E. BLEU

According to BLEU, the more *n*-gram matches with the reference translation, the better the candidate translation is. A modified precision score is calculated for the whole corpus by adding the clipped counts of matches (the total count of each candidate is clipped by the maximum number of times the word occurs in any single reference translation) and dividing the sum by the total number of *n*-grams in the candidate. The weighted average of the logarithm of the modified precisions accounts for the exponential decay in precision scores as *n*-grams get of higher order (1). The brevity penalty is computed over the entire corpus on best match reference lengths (2), where c denotes the candidate length, and r the best match reference length [15]. For calculating BLEU scores we use the NLTK script available at https://www.nltk.org/_modules/nltk/translate/bleu_score.html.

$$BLEU = brevityPenalty \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

$$brevityPenalty = \begin{cases} 1 & if\ c>r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c\leq r \end{cases} \quad (2)$$

F. CHRF++

CHRF uses character *n*-gram F-score, excluding spaces. It is calculated as in (3), where CHRP stands for the percentage of *n*-grams in the candidate translation which have a

[1] A decision tree provided to aid the annotation process can be found at http://www.qt21.eu/downloads/ annotatorsGuidelines-2014-06-11.pdf.

counterpart in the reference, while CHRR stands for the percentage of *n*-grams in the reference which are also present in the candidate. Parameter β assigns β times more importance to recall than to precision. It has been shown that the optimal option for β parameter is the value of 2, and for character *n*-grams the value of 6 [16]. CHRF++ score per segment is obtained by adding word 2-grams to the character 6-grams. We use the original python script for calculating the CHRF++ score, available at https://github.com/m-popovic/chrF.

$$\text{CHRF}\beta = (1+\beta)^2 \times \frac{\text{CHRP} \times \text{CHRR}}{\beta^2 \text{CHRP} + \text{CHRR}} \quad (3)$$

IV. Error Analysis

The most represented error issue in all three texts is mistranslation (Fig. 3). Remaining errors are mostly grammatical, e.g. case concordance, word order or incorrect function word.

An example of a *mistranslation* error is given in Fig. 4. An example of a sentence marked as illegible is given in Fig. 5. Although each error could be marked separately, according to MQM guidelines, the sentence has enough errors to be marked as illegible.

The translation of the manual obtained by Google Translate is surprisingly good. The most represented error category is again mistranslation. It is worth noting that certain parts are even translated into English instead of Croatian. This is probably due to the fact that many terms do not have a standardized equivalent or a standardized equivalent has not entered popular use so the English alternative is acceptable.

Fig. 6 is an example of the news sentence translated by Bing which exhibits several error issues, such as mistranslation, incorrect word form, i.e. tense under the grammar subcategory, and word order issue.

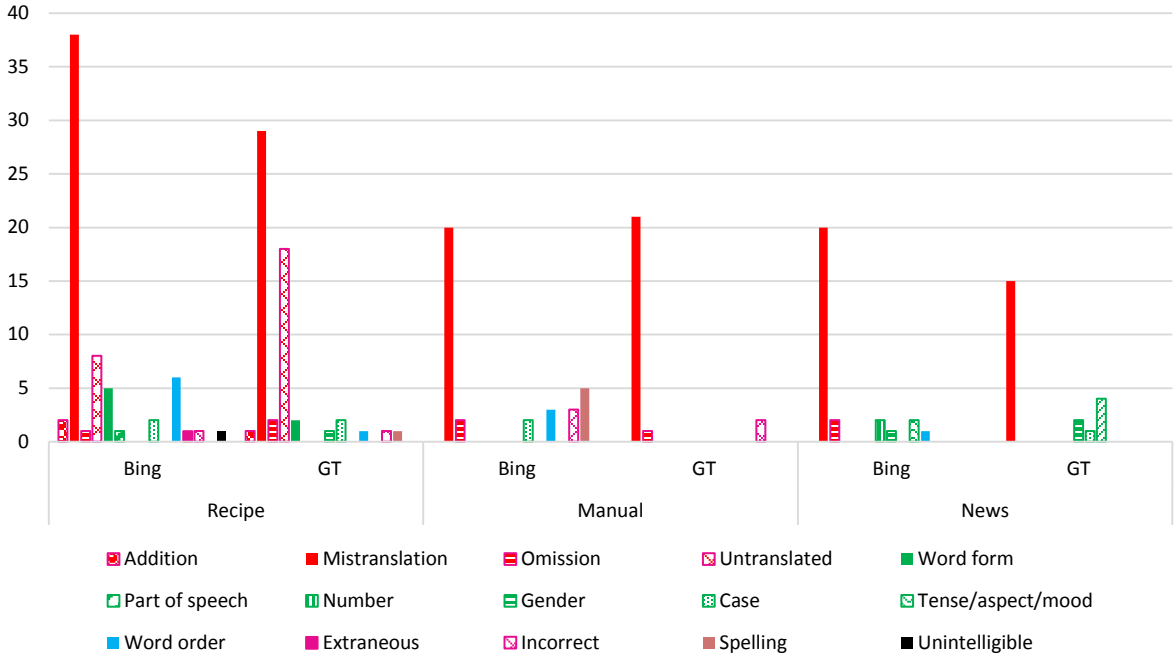

Fig. 3. MQM Issues Found.

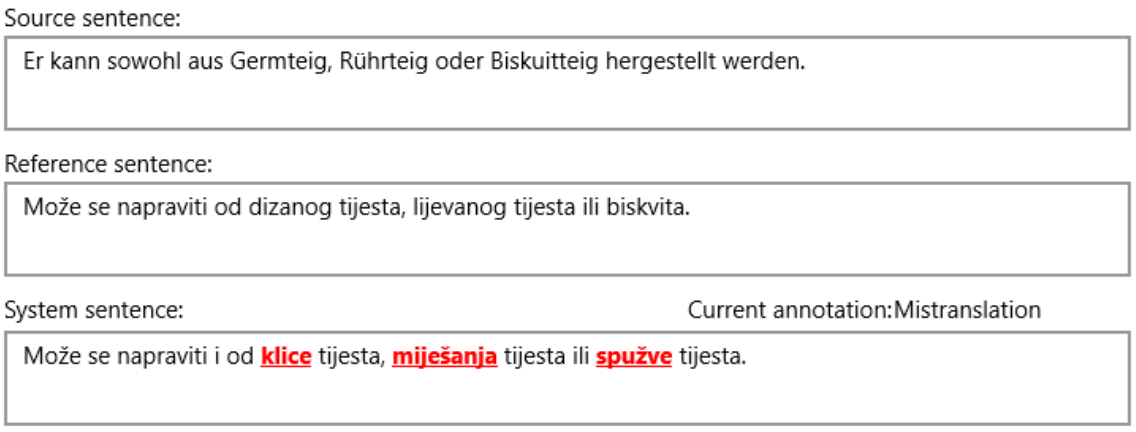

Fig. 4. Mistranslation Error in the Recipe Translated by Bing.

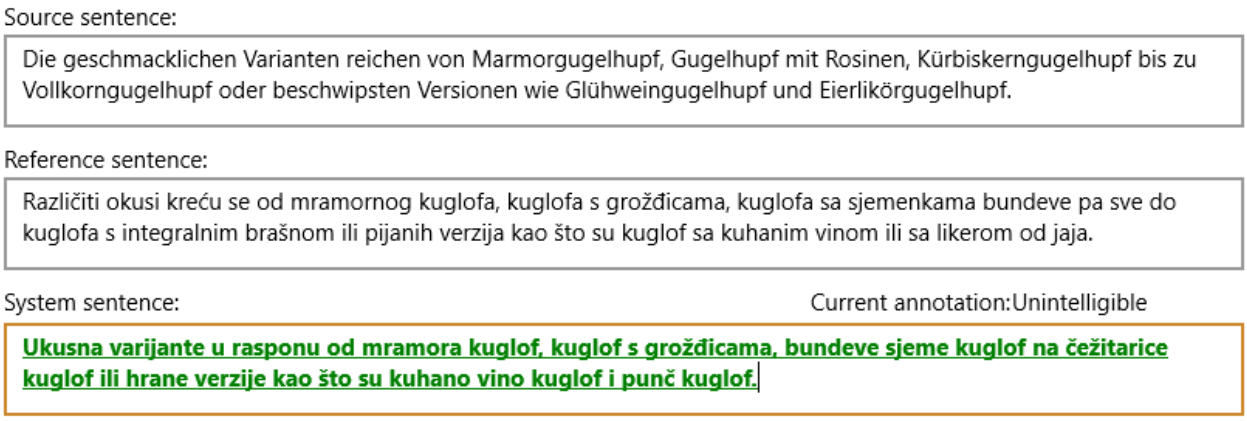

Fig. 5. Illegible Sentence in the Recipe Translated by Bing.

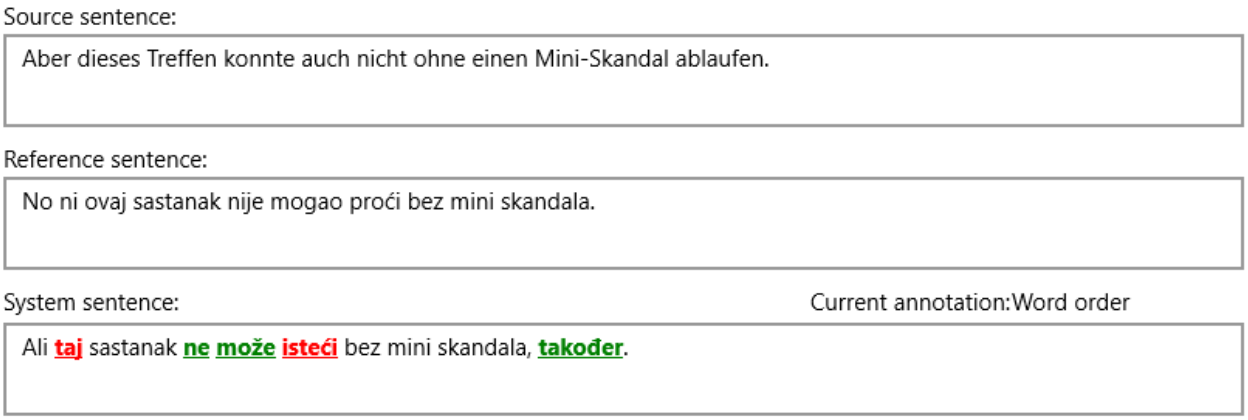

Fig. 6. A Sentence with Several Issues in the News Text Translated by Bing.

V. Results

BLEU and CHRF++ scores, as well as the average number of errors per sentence and the total number of errors per text are reported in Table II. Error counts per each category are presented in Fig. 3. For the sake of clarity, when possible, errors of the same superordinate category are shown in the same color with differing patterns. This being said, accuracy errors are presented in red, grammatical errors belonging to a word form subcategory in green, and grammatical errors relating to function words in fuchsia.

It is worth noting that GT manages to translate quite different number of sentences per each text flawlessly– 6% in the recipe, 21% in the manual, and 63% in the news, which indicates its varying performance on different text types. Two variants of CHRF++ score are reported–overall document level (F2), and macro averaged document level F-score (avgF2), which is the arithmetic average of sentence level scores.

The Pearson coefficient is used to calculate segment-level correlations between automatic metrics and total error counts (Table III), and between automatic metrics and two most represented error categories (Table IV). The Pearson coefficient ranges from +1 to −1, where +1 expresses total positive correlation, i.e. by increasing error counts, automatic metric scores increase, and −1 total negative correlation, i.e. by increasing error counts, automatic metrics scores decrease. A value of 0 denotes that there is no linear correlation between the inspected variables.

TABLE II. The Results of Automatic and Human Evaluation of MT

		Text					
		Recipe		***Manual***		***News***	
		Bing	***GT***	***Bing***	***GT***	***Bing***	***GT***
BLEU		53.81	52.93	39.49	38.55	47.72	46.79
CHRF++	F2	45.15	47.36	54.85	61.95	62.81	66.47
	avgF2	43.82	47.05	53.11	58.66	63.46	67.50
Error analysis	avg	3.66	3.22	1.59	1.09	2	1.57
	total	66	58	35	24	28	22

TABLE III. Segment-Level Correlation between Automatic Metric Scores and Total Number of Errors

	Text					
	Recipe		***Manual***		***News***	
	Bing	***GT***	***Bing***	***GT***	***Bing***	***GT***
BLEU	0.0016	0.3822	0.3462	0.2126	-0.0011	0.2447
CHRF++	-0.2507	-0.5004*	-0.0478	-0.3954	-0.6872**	-0.4671

* statistically significant at 5%
** statistically significant at 1%

TABLE IV. SEGMENT-LEVEL CORRELATION BETWEEN AUTOMATIC METRIC SCORES ON ONE HAND AND ACCURACY CATEGORY AND MISTRANSLATION SUBCATEGORY ON THE OTHER HAND

		Text					
		Recipe		***Manual***		***News***	
		Bing	***GT***	***Bing***	***GT***	***Bing***	***GT***
BLEU	acc	0.0359	0.4342	0.1239	0.3109	0.1882	0.1432
	mis	0.1087	0.3780	0.1231	0.3341	0.2329	0.1432
CHRF++	acc	-0.2752	-0.5404*	-0.2180	-0.4074	-0.5131	-0.3909
	mis	-0.2311	-0.5661*	-0.1438	-0.4017	-0.4490	-0.3909

VI. DISCUSSION

As far as the correlation is concerned, segment-level BLEU shows a positive correlation with human judgments (Table III, Table IV). This means that the BLEU score increases with the increase of errors, as if it is an error metric and not a precision-based one. Translations abundant with errors should be scored lower. A negative relationship exists between CHRF++ and human judgments. However, hardly any correlation proves to be statistically significant. Those rare which are significant are related to CHRF++. The abbreviations *acc* and *mis*, used in Table IV, refer to accuracy and mistranslation, respectively.

The examined automatic metrics do not even agree on the ranking of MT engines for the selected genres (Table II). According to BLEU, the most suitable genre for the selected MT engines is the one regarding recipes, followed by news and manual. On the other hand, CHRF++ rates news the best, followed by manual, and lastly recipe. The human annotator attributes the highest number of errors to the recipe, in line with CHRF++, while the other two genres are pretty close regarding the number of errors. However, CHRF++, unlike BLEU, manages to rank them correctly. If taking only GT into consideration, then the second best scoring translation according to the human annotator is the manual, while the best scoring is the news text. While the difference between the systems in terms of BLEU is less than one point, in terms of CHRF++ it ranges from 2 to over 7 points.

Although the authors in [20] show that BLEU best correlates with human judgments in domains containing short and simple sentences, but is surpassed by CHRF in cases with more complex syntactic structures and longer contexts, our evaluation gives advantage to CHRF++ in all three genres. We cannot make any conclusions on the effects of low-quality translations on the correlation since hardly any correlation proves to be statistically significant. Although not even human translations would obtain a score of 1 due to high variability of translations, having multiple references could increase the overall BLEU score and some segment-level scores and affect the correlation results presented in this paper.

VII. CONCLUSION AND FUTURE WORK

The conducted evaluation is of a black box type. Three texts of different genre are selected in order to examine translations produced by two popular MT services in the German-Croatian language direction.

Automatic metrics employed in the paper depend on the availability of human reference translations. BLEU is chosen, despite of its many drawbacks, as it is the *de facto* standard in MT community, and CHRF++ is chosen as it has potential in dealing with morphologically rich target languages.

MQM Slavic tagset compliant error analysis of translations is performed in TREAT by one annotator who is a native of Croatian and has a BA in the German language. The training is conducted prior to annotation in order to familiarize the annotator with MQM and TREAT.

Pearson correlation coefficients are calculated to check how close automatic evaluation reflects manual judgments. CHRF++ metric, which works on character level and which is enhanced with word *n*-grams, proves to correlate better with human judgments than BLEU, which is a metric that works only on word level. This is valid for all three genres.

A major drawback of this study, beside the fact that only one reference translation is present, is the inevitable trade-off made between the size of the sample and the integrity of the results. Manual error analysis performed in this study is extremely expensive and time consuming. Automatic metrics are, on the other hand, quick and cheap to use. However, they are of no use if they do not correlate with human judgments.

A pairwise comparison of metrics within other classes and between classes is purposefully excluded from this study and left for future work for the sake of clarity and conciseness. Besides enlarging the evaluation set, involving more annotators would be considered beneficial, despite the usual low inter-annotator agreement in such tasks. A particular focus of our future work will be put on the results of neural network-based metrics. Since hardly any statistically significant correlation is detected, neither on the total number of errors, nor on the most prominent error category and subcategory, our future work will also entail a weighing scheme which will give weights to different error issues in the total error counts. Correlation between automatic metrics and a scoring mechanism provided with MQM, which includes weights on a four-level scale, will also be investigated.

REFERENCES

[1] Way, "Quality Expectations of Machine Translation," Springer, Cham, pp. 159–178, 2018.

[2] M. Sepesy Maučec and G. Donaj, "Machine Translation and the Evaluation of Its Quality," in Natural Language Processing-New Approaches and Recent Applications, 2019.

[3] A. Secara, "Translation Evaluation - a State of the Art Survey," in Proceedings of the eCoLoRe/MeLLANGE Workshop, 2005, pp. 39–44.

[4] E. Chatzikoumi, "How to evaluate machine translation: A review of automated and human metrics," Nat. Lang. Eng., pp. 1–25, 2019.

[5] S. Stymne, "Pre- and Postprocessing for Statistical Machine Translation into Germanic Languages," in Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Student Sess., no. June, 2011, pp. 12–17.

[6] S. Stymne and L. Ahrenberg, "On the practice of error analysis for machine translation evaluation," in 8th Int. Conf. Lang. Resour. Eval., 2012, pp. 1785–1790.

[7] M. Popovic and A. Burchardt, "From Human to Automatic Error Classification for Machine Translation Output," in Proceedings of the 15th International Conference of the European Association for Machine Translation, no. May, 2011.

[8] A. Lommel, H. Uszkoreit, and A. Burchardt, "Multidimensional Quality Metrics : A Flexible System for Assessing Translation Quality," vol. 12, no. Tradumàtica 12, pp. 455-463., 2014.

[9] F. Klubička, A. Toral, and V. M. Sánchez-Cartagenac, "Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation," Prague Bull. Math. Linguist., no. 108, pp. 121–132, 2017.

[10] S. Majcunic, M. Matetic, and M. Brkic Bakaric, "Translation Error Analysis in TREAT : A Windows App Using the MQM Framework," Zb. Veleučilišta u Rijeci, vol. 7, no. 1, pp. 149–162, 2019.

[11] M. Fomicheva and L. Specia, "Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments," Comput. Linguist., vol. 45, no. 3, pp. 515–558, 2019.

[12] O. Bojar, Y. Graham, and A. Kamran, "Results of the WMT17 Metrics Shared Task," in Proceedings of the Conference on Machine Translation (WMT), 2017, vol. 2, pp. 293–301.

[13] O. Abend and A. Rappoport, "Universal conceptual cognitive annotation (UCCA)," in ACL 2013 - 51st Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., vol. 1, 2013, pp. 228–238.

[14] Q. Ma, J. Wei, O. Bojar, and Y. Graham, "Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges," in Proceedings of the Fourth Conference on Machine Translation, 2019, vol. 2, no. Day 1, pp. 62–90.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "B LEU : a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, no. July, pp. 311–318.

[16] M. Popovic, "CHRF : character n-gram F-score for automatic MT evaluation," in Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, no. September, pp. 392–395.

[17] M. Popovic, "CHRF++ : words helping character n-grams," in Proceedings ofthe Conference on Machine Translation, 2017, vol. 2, no. 1, pp. 612–618.

[18] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, 2002, pp. 138–145.

[19] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in AMTA 2006 - Proc. 7th Conf. Assoc. Mach. Transl. Am. Visions Futur. Mach. Transl., 2006, pp. 223–231.

[20] M. Chinea-Rios, A. Peris, and F. Casacuberta, "Are Automatic Metrics Robust and Reliable in Specific Machine Translation Tasks?," in 21st Annual Conference of the European Association for Machine Translation, 2018, no. May, pp. 89–98.

[21] M. Brkic Bakaric, N. Babic, L. Dajak, and M. Manojlovic, "A comparative error analysis of English and German MT from and into Croatian," in InFuture 2017, 2017, pp. 31–41.

[22] R. Fiederer and S. O'Brien, "Quality and machine translation: A realistic objective?," J. Spec. Transl., no. 11, pp. 52–74, 2009.