Evaluation of Various Approaches to Compute BLEU Metrics

Lucia Benkova and Ľubomír Benko

Constantine the Philosopher University in Nitra
Nitra 94901, Slovakia {lucia.benkova,lbenko}@ukf.sk

Abstract. Evaluation of machine translation (MT) performance, as the concept of quality, is closely related to the concept of optimization. Over recent decades, several approaches to evaluate MT quality have been proposed. Each approach brings new metrics for MT evaluation, and/or MT performance. The aim of our study is to show which of metrics based on precision that have been proposed so far are suitable for evaluating the quality of translation from English to Slovak in the domain of journalistic texts. We focus on the BLEU metric and its different variants that are available in the nltk libraries and the Python library. We attempt to determine which of the examined variants of the BLEU metric are redundant. The results of our research show the redundancy of BLEU-1 metric variants from the PyTorch library with respect to the newspaper style and neural MT. On the contrary, a statistically significant difference was shown by the PoSBLEU-1+1 and nltk-based BLEU-1 variants.

Keywords: Neural machine translation, Statistical machine translation, Automatic evaluation, BLEU, Slovak language, Text analysis

1 Introduction

The paper offers an evaluation of different approaches to automatic metrics for the evaluation of machine translation suitable for the Slovak language. In our previous research [1], [2], we used standard error rate and accuracy metrics such as PER, WER, TER, and BLEU. In this paper, we focus only on the state-of-art metric of accuracy, namely the BLEU-n metric from which we expect relevant results for the Slovak language and which offers open-source access to the source data and metric parameters. The BLEU metric is widely used to measure the quality of machine translation. We focus on freely Google Translate service as one of the most used online neural MT systems today.

The rest of the paper is structured as follows. The following section describes the related work of automatic metrics for MT evaluation. The third section focuses on the dataset description and methods used in the experiment. The fourth section deals with the results of the experiment where we compare various approaches to an accuracy automatic metric. The last section provides the conclusion and future work.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 71–78, 2022.

2 Related Work

Basic error-rate metrics include PER [3], WER [4] and TER [5] operating on the calculation of edit distance, the so-called Levenshtein distance, i.e., which provides the minimum number of edit operations (insertions, deletions or substitutions) needed to match two sequences of words. The aforementioned metrics differ from each other in their relation to word order, word position in the sentence, and translation penalty. Among the most common accuracy metrics is the BLEU-n metric [6], which, despite several flaws, is still very popular and standard within the users. BLEU-n is based on the geometric mean of the n-grams precision of length 1 to 4 and a penalty of sentence shortness (brevity penalty).

Many authors focus their research around the BLEU metrics and its variations. Benkova et al. [1] focus on the comparison of phrase-based statistical MT systems (Google SMT and mt@ec) and neural MT systems (Google NMT and eTranslation) using automatic metrics for MT evaluation from English to Slovak. The research was conducted using residuals to compare the scores of BLEU-n metrics. The results confirm the assumption of better neural MT quality regardless of the system used. Statistically significant differences between the SMT and NMT were found in favour of NMT based on all BLEU-n scores. Munkova et al. [2] focused on an evaluation of automatic measures of error rate and accuracy when validating the quality of MT output from the synthetic Slovak language to the analytical English language. They used multiple comparisons for the analysis and icon graphs to visualize the results. The results showed that all examined metrics, which are based on textual similarity, except the f-measure, are needed to be included in MT quality evaluation when analyzing MT output based on sentence. The authors [7] presented a deep evaluation and error analysis of five paraphrase generation modules of the Watson project. The results revealed the most problematic sources of errors in the generation process and helped with further improvements to the system.

Biesialska et al. [8] analysed the performance of the statistical and neural approaches to MT. They compared phrase- 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). They applied back translation into their MT system and examined the scores of BLEU-n score [6]. The results showed that for the Czech–Polish language pair, the BLEU score was relatively low, which was explained by the language distance.

Almahasees [9] focused on the comparison of two MT systems, Google Translate and Microsoft Bing translator. Both systems were based on an SMT system for the English-Arabic language pair. The comparison of the MT outputs of journalistic texts was conducted using the standard automatic evaluation metric BLEU-n. The results were in favour of Google Translate, where Bing generated semantically different sentences.

3 Materials and methods

The aim of the research is to filter out the redundant metrics of automatic MT evaluation. This study can later serve as a reference to identify redundant metrics from various sets of similar metrics (BLEU, ROUGEs metrics or other metrics of error rate or accuracy).

3.1 Dataset composition

We used the dataset which consists of 66 original English journalistic texts (39 354 word tokens). These texts were translated by Google Translate using SMT and NMT. Besides, texts were also translated by two professional human translators (HT) and post-edited by another professional human translator (PEMT) using our online system OSTPERE (Online System for Translation, Post-Editing, Revision, and Evaluation) [10], [11]. The translation direction was from English to Slovak, as Slovak is one of the official EU languages and contains an inflected morphology and loose word order [12]. The table 1 gives a summary of the composition of the dataset.

Table 1: Lexico-grammatical dataset composition.

Feature type	**Feature name**	**SMT**	**NMT**	**HT**	**PEMT**	**SRC**
Readability	Average sentence length	17.164	17.236	17.880	17.994	19.414
	Average word length	5.571	5.664	5.764	5.706	4.951
	Number of short sentences	487	493	466	449	413
	Number of long sentences	1557	1551	1578	1595	1631
Lexico-grammatical	Frequency of noun	9314	9365	9999	9877	8713
	Frequency of adjective	4436	4407	4659	4801	3213
	Frequency of verb	4218	4400	4437	4389	5246
	Frequency of determiner	1918	1876	1973	1971	3953
	Frequency of adposition	3735	3875	4129	4155	4680
	Frequency of proper noun	2231	2198	2165	2195	3411
	Frequency of coordinating conj.	1338	1311	1396	1334	1246
	Frequency of subordinating conj.	1352	1403	1281	1377	853
	Frequency of interjection	18	8	9	10	15
	Frequency of adverb	1307	1247	1339	1382	1653
	Frequency of pronoun	1055	1260	1417	1324	2615
	Frequency of auxiliary	1626	1299	1257	1374	2432
	Frequency of numeral	1260	1311	1195	1302	1009
	Frequency of particle	573	598	777	764	1312
	Frequency of punctuation	6668	6674	6460	6646	5370
	Frequency of other	597	561	589	511	3

3.2 Methodology

The experiment is focused on the most popular metric of accuracy- BLEU. We have taken various libraries and approaches to calculate the BLEU metrics.

The BLEU metric [6] is considered a state-of-art automatic evaluation metric. The metric is based on the geometric mean of n-gram precisions and brevity penalty (a length-based penalty). BLEU performs well at the corpus level but lags significantly at the sentence level. Lin and Och [13] applied various smoothing techniques to BLEU to obtain better results at the sentence level. Suppose we have similar n-grams for $n = 1...N$ (often $N = 4$). Let m_n be the original number of hits and m'_n be the number of hits of the modified n-gram. One smoothing technique says that if the number of matching n-grams is equal to 0, then we use a small positive value ε to replace 0 for n in the range from 1 to N.

$$m'_n = \varepsilon, if\ m_n = 0.$$

There are seven smoothing techniques that are used mainly to evaluate the output based on sentences. We have focused on the second smoothing technique (the other technique's results did not yield relevant scores) that adds 1 to the number of matching n-grams and the total number of n-grams for n in the range from 2 to N.

$$l'_n = l_n + 1, for\ n\ in 2..N.$$

A different approach to evaluating machine translation is offered by the PoS-BLEU metric [14]. It is one of the metrics focusing on the syntactic structure of the translation output, where PoS tags are the input of the calculation instead of words.

In this experiment we will focus on the BLEU-1 metric and its variations (nltk and PyTorch library, with and without smoothing function, PoSBLEU-1+1). We expect that there will be no differences between the various BLEU-1 metrics approaches and therefore it will not play a role which approach we use in machine translation evaluation. The methodology of the experiment consists of the following steps:

1. obtaining the unstructured text data (source text) and removing the document formatting,
2. machine translation using various systems (SMT, NMT)
3. human translation of the documents,
4. post-editing of the machine translation,
5. segment alignment between the source text, machine translations, human translation and post-edited text,
6. human evaluation of examined machine translation based on model [15],
7. automatic evaluation of examined machine translation using various metrics (BLEU-1 for this experiment), where as reference text were chosen as human translation so post-edited text,
8. comparison of the translation quality based on the accuracy and translation system (SMT, NMT),
9. evaluation of obtained results.

4 Results

We have focused to identify the redundancy between various approaches to the BLEU-1 metric. We have used Python-based libraries to implement the BLEU metric. We used the library nltk, PyTorch (with and without the smoothing function) and our own function to obtain the results of POSBLEU. The POSBLEU metric needed a morphological annotation of texts, so we used the Stanza library which contains a model for the Slovak language. We have analysed the texts translated by SMT and NMT separately. Both outputs were evaluated by a human and for the SMT were identified 1574 segments that contained an error and only 470 segments were evaluated as correct. In the case of NMT, 1658 segments were correct and only 386 contained an error.

To test the global null hypotheses, we used adjusted tests for repeated measurements (Huynh-Feldt adjustment), due to the violation of the sphericity condition of the covariance matrix. If the covariance matrix sphericity condition is not satisfied, the magnitude of the type I. error increases. The epsilon represents the degree of violation of the sphericity condition. An epsilon equal to one represents the satisfaction of the condition. Conversely, the smaller it is, the more the sphericity condition is violated.

When testing the global null hypotheses, epsilon values were less than one (Table 2). In the case of SMT, null hypotheses are rejected with 99.9% confidence (at the 0.001 significance level). The hypotheses assert that group segment accuracy does not depend on variations in BLEU-1 accuracy metrics and combinations of BLEU-1 and segment accuracy factors (manual evaluation 0/1).

Similarly, in the case of the NMT, it has been shown that the accuracy of the segments studied depends on the variation of the BLEU-1 accuracy metrics. In contrast, the dependence on the combination of BLEU-1 and segment accuracy factors (manual evaluation 0/1) was not confirmed.

In terms of multiple comparisons (Table 3), we have identified three homogeneous groups (**** - $p > 0.05$) in the degree of accuracy of the examined segments. A statistically significant difference in segment accuracy rates was demonstrated between POSBLEU_1+1 and the others, and similarly between

Table 2: Huynh-Feldt adjustment for BLEU-1 and segment accuracy for (a) SMT and (b) NMT.

(a) **NMT=0**	**H-F Epsilon**	**H-F Adj. df1**	**H-F Adj. df2**	**H-F Adj. p**
BLEU-1	0.5087	1.5260	3116.1310	0.0000
BLEU-1*Evaluation_Error	0.5087	1.5260	3116.1310	0.000
(b) **NMT=1**	**H-F Epsilon**	**H-F Adj. df1**	**H-F Adj. df2**	**H-F Adj. p**
BLEU-1	0.5558	1.6675	3404.9990	0.0000
BLEU-1*Evaluation_Error	0.5558	1.6675	3404.9990	0.8424

Table 3: Multiple comparisons for various BLEU-1 metrics and segment accuracy for (a) SMT and (b) NMT.

(a) NMT=0 **BLEU-1**	**Mean**	**1**	**2**	**3**
PyTorch_BLEU-1_smooth	0.504	****		
PyTorch_BLEU-2	0.504	****		
BLEU-1	0.626		****	
POSBLEU-1+1	0.719			****
(b) NMT=1 **BLEU-1**	**Mean**	**1**	**2**	**3**
PyTorch_BLEU-1_smooth	0.519	****		
PyTorch_BLEU-2	0.519	****		
BLEU-1	0.664		****	
POSBLEU-1+1	0.743			****

BLEU-1 and the other metrics ($p < 0.05$). On the other hand, a statistically significant difference was not identified between the PyTorch metrics. The results are the same for both translation systems, the expected higher accuracy rates were achieved for NMT. From this point of view, the redundant metric will be precisely one of these PyTorch metrics.

The results showed us that the PyTorch metrics are redundant. In this case, the smoothing function that was introduced to improve the evaluation based on segments did not produce different results than the corpus-based BLEU-1 metric from the PyTorch library. In the future, we can omit the smoothing function variant of the BLEU-1 metric.

5 Conclusion

The paper deals with the metrics of the automatic MT evaluation and is a basis for our future experiments. We have introduced a methodology to filter out the redundant metrics that were experimented on using the BLEU-1 metric. This will be expanded in future work that will deal with a greater number of automatic metrics, that will be grouped based on related characteristics. We would also like to compare newer metrics, like ChrF++ [16], BEER [17], LEPOR [18], COMET [19], with older like NIST [20], ROUGE [21], METEOR [22]. The aim is to select the most appropriate automatic metrics for evaluating MT output into Slovak. In this paper, we have shown that various approaches to calculate the BLEU-1 metric show significant differences. However, the use of the smoothing function does not produce significantly different results than using the corpus-based BLEU-1 metric.

Acknowledgement This work was supported by the Slovak Research and Development Agency under the contract No. APVV-18-0473 and by the projects UGA VII/2/2022 and UGA VII/1/2022.

References

1. Benkova, L., Munkova, D., Benko, Ľ., Munk, M.: Evaluation of English–Slovak Neural and Statistical Machine Translation. Applied Sciences. 11, (2021). https://doi.org/10.3390/app11072948.
2. Munkova, D., Hajek, P., Munk, M., Skalka, J.: Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy. Procedia Comput Sci. 171, 1327–1336 (2020). https://doi.org/10.1016/j.procs.2020.04.142.
3. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H.: Accelerated DP based search for statistical translation. In: European Conference on Speech Communication and Technology. pp. 2667–2670. Rhodes, Greece (1997).
4. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas. pp. 223–231 (2006).
5. Nießen, S., Och, F.J., Leusch, G., Ney, H.: An evaluation tool for machine translation: Fast evaluation for MT research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000). pp. 39–45 (2000).
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. , Philadelphia (2002).
7. Burgerová, V., Horák, A.: Evaluation and Error Analysis of Rule-based Paraphrase Generation for Czech. In: Horák, A., Rychlý, P., and Rambousek, A. (eds.) Proceedings of the Thirteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2019. pp. 33–39. Tribun EU, Brno (2019).
8. Biesialska, M., Guardia, L., Costa-jussa, M.R.: The TALP-UPC System for the WMT Similar Language Task: Statistical vs Neural Machine Translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp. 185–191. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-5424.
9. Almahasees, Z.M.: Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. International Journal of Languages, Literature and Linguistics. 3, 1–4 (2017). https://doi.org/10.18178/ijlll.2017.3.1.100.
10. 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. pp. 365–376. Springer Verlag (2020). https://doi.org/10.1007/978-3-030-11932-4_35.
11. Munková, D., Kapusta, J., Drlík, M.: System for Post-Editing and Automatic Error Classification of Machine Translation. In: DIVAI 2016: 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, May 2 – 4, 2016. pp. 571–579. Wolters Kluwer, ISSN 2464-7489, Sturovo (2016).
12. Kosta, P.: Targets, Theory and Methods of Slavic Generative Syntax: Minimalism, Negation and Clitics. In: Kempgen, Sebastian / Kosta, Peter / Berger, Tilman / Gutschmidt, Karl (eds.). Slavic Languages. Slavische Sprachen. An International Handbook of their Structure. In: Kempgen, S., Kosta, P., Berger, T., and Gutschmidt, K. (eds.) 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. pp. 282–316. Berlin, New York: Mouton. de Gruyter (2009).
13. Lin, C.-Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the

42nd Annual Meeting on Association for Computational Linguistics - ACL '04. pp. 605-es. Association for Computational Linguistics, Morristown, NJ, USA (2004). https://doi.org/10.3115/1218955.1219032.
14. Popović, M., Ney, H.: Syntax-oriented evaluation measures for machine translation output. In: Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09. p. 29. Association for Computational Linguistics, Morristown, NJ, USA (2009). https://doi.org/10.3115/1626431.1626435.
15. Vaňko, J.: Kategoriálny rámec pre analýzu chýb strojového prekladu. In: Munkova, D. and Vaňko, J. (eds.) Mýliť sa je ľudské (ale aj strojové). pp. 83–100. UKF v Nitre, Nitra (2017).
16. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Stroudsburg, PA, USA (2015). https://doi.org/10.18653/v1/W15-3049.
17. Stanojević, M., Sima'an, K.: Evaluating MT systems with BEER. The Prague Bulletin of Mathematical Linguistics. 104, 17–26 (2015). https://doi.org/10.1515/pralin-2015-0010.
18. Han, A.L.F., Wong, D.F., Chao, L.S.: LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. In: Proceedings of COLING 2012: Posters. pp. 441–450. The COLING 2012 Organizing Committee, Mumbai, India (2012).
19. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A Neural Framework for MT Evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Stroudsburg, PA, USA (2020). https://doi.org/10.18653/v1/2020.emnlp-main.213.
20. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. 138–145 (2002).
21. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004).
22. Lavie, A., Denkowski, M.: The Meteor metric for automatic evaluation of machine translation. Machine Translation. 23, 105–115 (2009).