Identification of Relevant and Redundant Automatic Metrics for MT Evaluation

Michal Munk[1], Daša Munková[2], and Ľubomír Benko[3(✉)]

[1] Department of Informatics, Faculty of Natural Sciences,
Constantine the Philosopher University in Nitra, Nitra, Slovak Republic
mmunk@ukf.sk
[2] Department of Translation Studies,
Constantine the Philosopher University in Nitra, Nitra, Slovak Republic
dmunkova@ukf.sk
[3] Institute of System Engineering and Informatics,
University of Pardubice, Pardubice, Czech Republic
lubomir.benko@gmail.com

Abstract. The paper is aimed at automatic metrics for translation quality assessment (TQA), specifically at machine translation (MT) output and the metrics for the evaluation of MT output (Precision, Recall, F-measure, BLEU, PER, WER and CDER). We examine their reliability and we determine the metrics which show decreasing reliability of the automatic evaluation of MT output. Besides the traditional measures (Cronbach's alpha and standardized alpha) we use entropy for assessing the reliability of the automatic metrics of MT output. The results were obtained on a dataset covering translation from a low resource language (SK) into English (EN). The main contribution consists of the identification of the redundant automatic MT evaluation metrics.

Keywords: Machine translation · Evaluation · Automatic metrics · Reliability · Entropy · Redundancy

1 Introduction

Machine translation (MT), its specifics and function are still relatively under researched fields, not only in Translation Studies, but also in Natural Language Processing. As its tools are still in their infancy and need a lot of adjustments and improvements to achieve better translation quality in target languages, MT theory itself is in its early stages in the context of low resource languages. This is caused not only by the short and recent development time of MT tools and MT systems, but also because machine translation is an interdisciplinary field comprising various research areas such as Translation studies, Computer engineering or Computational linguistics. Translation studies is focused on the evaluation of machine translation output from the point of view of linguistic phenomena, whereas Computer engineering or Computational linguistics are focused on the determination of the effectiveness of existing MT systems and on the optimization of algorithms implemented in MT systems as well as on the

C. Sombattheera et al. (Eds.): MIWAI 2016, LNAI 10053, pp. 141–152, 2016.
DOI: 10.1007/978-3-319-49397-8_12

performance of MT systems. Progress relies on translation quality assessment through systematic effective evaluation approaches. Better evaluation metrics lead to better machine translation [1]. There are many evaluation approaches used in MT evaluation. Babych et al. [2] examined the effectiveness of the performance of MT system translating from low-resource languages into English via closely-related and well-developed translation resources. Adly and Al Ansary [3] evaluated the effectiveness of MT system based on the Interlingua approach. Vandeghinste et al. [4] evaluated the METIS-II system, Machine Translation System for Low Resource Languages, based on the automatic metrics BLEU, NIST and TER.

In general, there are two main approaches to MT evaluation: Glass box and Black box. We focus on a black box approach to MT evaluation, measuring the performance of a system upon a same test set and within the black box on intrinsic metrics. Intrinsic metrics – manual and automatic - are used to assess the accuracy of MT output. They focus on the quality of MT output and they compare MT output with one or more references (high quality translation, usually done by a human translator). Manual (human) intrinsic metrics assess the quality of Mt output as fluency and adequacy by human, which is not only subjective, but expensive, slow and difficult to standardize. Vilar et al. [5] remarked that the subjectivity of manual evaluation causes a problem in terms of the lack of clear guidelines as to how to assign values to translations. Automatic intrinsic metrics correlate well with human judgements [6–10] using quantitative scores of adequacy and fluency [11]. They compute sentence similarity -matches based on comparisons between a set of references (fixed translations) and the corresponding MT output. Automatic evaluation offers easy, low cost and high speed of evaluation compared to human translation, which is regarded as the most reliable.

Statistical machine translation systems have been the most widely used for many recent surveys and events. Since 2006 the evaluation campaigns, during the Annual workshop on Statistical machine translation (WMT), have been organized by the special interest group of machine translation (SIGMT) focusing on European languages [12–21]. The tested language pairs are divided into two directions from English into other (French, German, Spanish, Czech, Hungarian, Haitian Creole and Russian) and vice versa, i.e. from other languages into English. These campaigns do not cover other European language – Slovak, which is the subject of this paper.

We demonstrate how the analysis of reliability of automatic intrinsic metrics using Cronbach's or Standardized alpha or entropy can help by the identification of the relevant and redundant intrinsic metrics of automatic MT evaluation. Moreover, it can be used as a starting point for the automatic identification of errors and the classification of MT output.

Issues of MT and MT evaluation are more topical since there are not many studies concerning the less resource languages, such as the inflectional Slovak language.

This paper is constructed as follow: Sect. 2 introduces seven automatic intrinsic metrics for MT evaluation, Sect. 3 describes experiment setting, Sect. 4 presents the results of the analyses and finally Sect. 5 consists of conclusions.

2 Intrinsic Automatic Metrics for MT Evaluation

Due to problems which manual/human intrinsic metrics deal with, automatic metrics have been widely used during MT evaluation campaigns. They compare MT output with human reference, which can comprise one single reference or multiple references for a single source sentence [22, 23].

Automatic evaluation of MT output can be conducted based on statistical principles (n-grams or edit distance), which means based on lexical similarities or on the use of deep linguistic structures (morphological, syntactic or semantic information), which means based on linguistic features.

In this paper we will focus only on automatic intrinsic metrics based on lexical similarity (n-grams which measure the overlap in word sequences and partial word order and also edit distance).

Precision, Recall and F-measure belong to standard and easy measures. Precision (P) and Recall (R) are based on the concordance of words in MT sentence (hypothesis) with the words in the reference, regardless of the position of the word in a sentence. They have a mutually inverse relationship, i.e. the higher precision, the lower recall and vice versa $Precision = \frac{correct\ words}{length\ hypothesis}$ and $Recall = \frac{correct\ words}{length\ reference}$.

F-measure is a combination of both, *Precision* and *Recall*. It originates in information retrieval and was adapted to machine translation. It is a weighted harmonic mean of precision and recall, $F - measure_{\alpha} = \frac{(1+\beta^2)*P*R}{R+\beta^2*P}$, where $\alpha, \beta \in R^+$ are parameters for the weights, whereby $\alpha = \frac{1}{1+\beta^2}$.

Bilingual Evaluation Understudy (BLEU) is a current standard and widely used metric for MT evaluation. It is a precision oriented metric, it is a geometric mean of *n*-gram i.e. it computes the number of n-grams in the MT output (hypothesis) which also occur in a reference (for n-gram of size 1-4 with the coefficient of brevity penalty).

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

where

$$BP = brevity\ penalty = \begin{cases} 1, & if\ hypothesis\ >\ reference \\ e^{1-\frac{reference}{hypothesis}}, & if\ hypothesis\ \leq\ reference \end{cases}$$

and

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

Remark 1. S means hypothesis sentence in the complete corpus C.

The *BLEU* metric reflects two linguistic phenomena of manual evaluation metrics-adequacy and fluency, i.e. to semantically correct words and to word order. Lin and Och [22] or Papineni et al. [6] proved that shorter *n*-grams correlates better with

adequacy with 1-gram being the best predictor, while longer *n*-grams has better fluency correlation.

Precision, Recall, F-measure and BLEU metrics are measures of accuracy based on lexical similarity. The other category of automatic intrinsic metrics of MT evaluation is a category based on edit distance. Metrics in this category are called metrics of error rates. They do not measure the concordance but they compute the minimum number of editing steps needed to transform MT output to reference.

Word Error Rate (WER) is based on the edit distance and takes into account the word order. Edit distance is the minimum number of edit operations like word insertions, substitutions and deletions necessary to transform the MT output into the reference. The number of edit operations is divided by the number of words in the reference. When multiple reference translations are given, the reported error for a translation hypothesis is the minimum error over all references

$WER(h,r) = \frac{min_{e \in E(h,r)}(insertion(e) + deletion(e) + substitution(e))}{|r|}$, where *insertion* (*e*) – number of adding words, *deletion* (*e*) – number of dropping words, *substitution* (*e*) – number of replacements (in sequence or path *e*), *r* is a reference of MT output *h* and $min_{e \in E(h,r)}$ is a minimal sequence of adding, dropping and replaced words necessary to transform the MT output (*h*) into the reference *r*.

Diversity of language expressions causes the existence of many correct translations even though they are marked as "wrong" order or "wrong" word choice by WER to the references.

Position-independent Error Rate (PER) is a solution for this problem. It does not take into account word order when matching MT output and reference [24]. It is similar to the recall measure. They use the same denominator. It computes the matches between words appearing in MT output (hypothesis) and in reference regardless of the word order in both sentences. It considers the reference and hypothesis as bags of words.

It takes into account excess words that are considered defective and should be removed for translations which are too long

$PER = 1 - \frac{correct - \max(0, length\ hypothesis - length\ reference)}{length\ reference}$.

Cover Disjoint Error Rate (CDER) is based on the Levenshtein distance. It uses the fact that the number of blocks in a sentence is the same as the number of gaps between them plus one. It does not add blocks movement to its calculation, it expresses that as a long jump operation (jump over the gaps between two blocks) and it does not penalize the transfer of entire blocks. *Long jump* is combined with the other steps of editing (*insertion*, *substitution* or *deletion*) and with the null operation in case of identity. In other words, it permits reordering of the whole blocks without penalization. Single words in the reference must be covered only once, while in the hypothesis they can be covered zero, one or more times

$CDER(h,r) = \frac{min_{e \in E(h,r)}(insertion(e) + deletion(e) + substitution(e) + longjump(e))}{|r|}$, where *insertion* (*e*) – number of adding words, *deletion* (*e*) – number of dropping words, *substitution* (*e*) – number of replacements (in sequence or path *e*) and *long jump* (*e*) – number of

long jumps, r is reference translation of hypothesis h and $min_{e \in E(h,r)}$ is minimal sequence of adding, dropping and replaced words necessary to transform the MT output (h) into the reference r.

3 Method

In this experiment we examined the performance of the MT system – Google translation Api (GT), which is a free web translation service offering translation from/to Slovak. We used automatic intrinsic metrics of MT evaluation, namely- metrics of accuracy (precision, recall, F-measure and Bleu-n) and metrics of error rate (WER, PER and CDER), which are described in depth in the previous section. We assessed the translation quality from morphologically complex language into analytical. In other words, we evaluated the translation quality from European language- Slovak into English. Primarily Slovak language is very rich in inflectional and derivational forms contrary to English with its limited morphological system. Also Slovak has a loose word order compared to English which has a fixed word order. We chose this translation direction for better scores from the metrics WER and BLEU.

We developed a dataset consisting of 360 sentences derived from an original text written without using a control language. Sentences were translated using the above mentioned MT system. We chose only 360 sentences, because the experiment was limited by time (it was a part of another experiment focusing on human translation and post-editing of MT output). Human translators had only 90 min to translate the text into English and to provide a reference translation for MT evaluation. By the text representation we arose from the transaction-sequence model which is further described in [25–27]. We used our system of automatic MT evaluation, in which all algorithms representing the measures were implemented, to obtain scores of the examined metrics. As explained above, all metrics calculate the scores based on the comparison of MT output with one human reference. For sentence alignment the algorithm and software Hunalign was used [28]. Hunalign is an algorithm combining length-based [29, 30] and dictionary (translation) based [31–34] approaches for corpus alignment at the sentence level. For better performance, the algorithm does not take into account the possibility of more than two sentences matching into one sentence.

The first objective of the research was to determine the reliability of automatic intrinsic metrics for MT evaluation using the analysis of reliability and entropy. These metrics can be divided into metrics of error rate (the higher values of these metrics, the lower the translation quality) and metrics of accuracy. The second research objective targeted the identification of relevant and reductant automatic intrinsic metrics for MT evaluation. We tried to identify which metrics decrease the total score of reliability of automatic MT evaluation of machine translation from Slovak to English.

Entropy can be described as a measure of the expected content of the information or uncertainty probability distribution. It is also described as the degree of disorder or randomness in a system. Based on Shannon's definition [35, 36], given a class random variable C with a discrete probability distribution $\{p_i = Pr[C = c_i]\}_{i=1}^{k}, \sum_{i=1}^{k} p_i = 1$

where c_i is the i^{th} class. Then the entropy $H(C)$ is defined as $H(C) = -\sum_{i=1}^{k} p_i \log p_i$, while the function decreases from infinity to zero and p_i takes values from interval 0–1 [35, 36].

4 Results

The analysis results showed that the examined automatic intrinsic metrics of error rate are considered highly reliable based on the direct estimation of reliability. As it is shown in Table 1, each metric correlates with the total score of the evaluation (*Avg inter-metrics correlation*: 0.885) and after their elimination the coefficient of reliability has not increased (*Cronbach's alpha*: 0.950; *Standardized alpha*: 0.953) except for the metric referring to word order (WER). After elimination of the metric *WER*, the coefficient of reliability- *Cronbach's alpha* increased from 0.947 to 0.964, which is insignificant. However, the metric WER is the most deviated from the others in the translation quality assessment.

Table 1. Statistics of automatic intrinsic metrics of error rate.

	Metrics-total correlation	Alpha if deleted	Metrics-total accuracy entropy
PER	0.878	0.934	0.934
WER	0.845	0.964	0.852
CDER	0.958	0.869	0.895

For the *entropy* calculation (Table 1), in the case of the analysis of automatic metrics characterizing the error rate of MT evaluation, individual metrics in comparison over accuracy metrics were used. *Entropy* was calculated for each sentence analysed using the specific metrics and for the comparison the average entropy of all sentences was used. From the definition [35] if the *entropy* is closer to 1, then the system is more irregular. The results of the *entropy* for each of the error rate metric correspond with the coefficient of reliability- *Cronbach's alpha*.

The same was shown by the metrics of accuracy. Based on the direct estimation of reliability, metrics precision, recall, F-measure and Bleu-n are considered highly reliable.

Each metric (Table 2) correlates (*Avg inter-metrics correlation*: 0.882) with the total score of evaluation and after their elimination, the *coefficient of reliability* has not increased (*Cronbach's alpha*: 0.975; *Standardized alpha*: 0.975) except for the metric *BLEU-4*. After the elimination of metric *BLEU*-4, the *coefficient of reliability- Cronbach's alpha* increased from 0.974 to 0.976, which is also insignificant (metric *BLEU-4* measures a score of sequence of four words including articles and prepositions).

After the first analysis concerning the reliability of metrics representing the error rate of MT output, we assumed, that the metrics Bleu-n would copy or behave like the

Table 2. Statistics of automatic intrinsic metrics of accuracy.

	Metrics-total correlation	Alpha if deleted	Metrics-total error rates entropy
Precision	0.854	0.973	0.832
Recall	0.939	0.967	0.859
F-measure	0.949	0.966	0.852
BLEU_1	0.933	0.967	0.855
BLEU_2	0.943	0.967	0.852
BLEU_3	0.900	0.970	0.763
BLEU_4	0.807	0.976	0.675

metric WER. This resulted from the fact (as we mentioned in the Sect. 2), that both measures refer to the syntactical structure of the sentence, namely to word order.

The estimations of the entropy of automatic metrics of accuracy (Table 2) were similarly calculated as in the case of the metrics of error rate. Also in this case, the average entropy of all sentences for each metric were used and the results relate with the coefficient of reliability- *Cronbach's alpha* with negligible variations. In case of *entropy*, it also showed that the metric *BLEU-4* deviates the most from the other metrics.

Based on the adjusted univariate test for repeated measures, the zero hypothesis reasoning that the score of automatic intrinsic measures of MT evaluation (PER, WER and CDER) does not depend on individual metrics of the error rate, is rejected at the 1 % significance level (*G-G Epsilon* = 0.6788, *G-G Adj. p* = 0.0000). The strictest metric of error rate was identified *WER* (approximately 63 %) and the loosest *PER* (approximately 47 %).

Based on the results of multiple comparisons- Tukey test (Table 3) three homogenous groups (*PER*), (*CDER*) and (*WER*) were identified in terms of the score of the automatic evaluation of MT. Statistically significant differences in the score between *PER*/*CDER*/*WER* and others were proved at the 5 % significance level.

Table 3. Homogeneous groups for automatic intrinsic metrics of error rate.

Metrics of error rate	Mean	1	2	3
PER	47.10	****		
CDER	56.98		****	
WER	63.06			****

Plot (Fig. 1) visualizes the differences between examined metrics. The means with error plot depicts the means and confidence intervals of metrics of error rate.

The metrics of error rate have a significant impact on the quality of MT evaluation, as well as that metrics PER, WER and CDER are relevant for automatic evaluation of MT output.

The second part of the analysis is similar to the previous, and differs only in the metrics. Based on the results of the adjusted univariate test for repeated measure

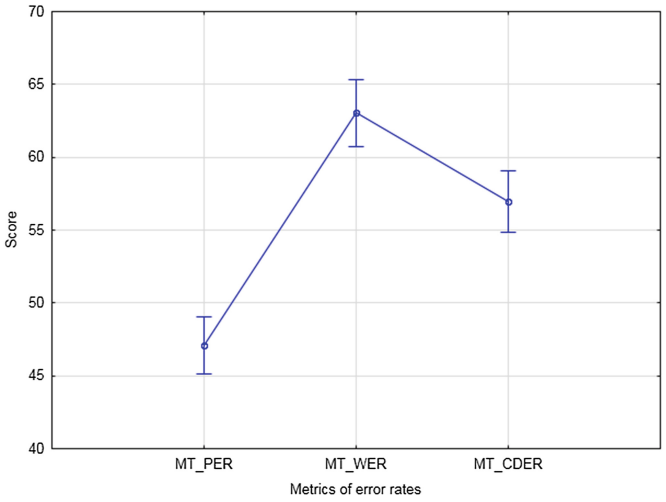

Fig. 1. The means with error plot for intrinsic metrics of error rates.

(*G-G Epsilon* = 0.3426, *G-G Adj. p* = 0.0000) the zero hypothesis reasoning that the score of automatic evaluation of MT does not depend on individual metrics of accuracy, is rejected at the 1 % significance level. The strictest metrics of accuracy (Table 4) were identified *BLEU-4*, *BLEU-3* and *BLEU-2* (approximately 14 %–32 %), and the loosest *Recall*, *BLEU-1*, *F-measure* and *Precision* (approximately 57 %–61 %).

Table 4. Homogeneous groups for automatic intrinsic metrics of accuracy.

Metrics of accuracy	Mean	1	2	3	4	5
BLEU-4	14.03		****			
BLEU-3	20.64			****		
BLEU-2	31.93				****	
Recall	56.86	****				
BLEU-1	57.50	****				
F-measure	58.25	****				
Precision	60.92					****

From multiple comparisons- Tukey test (Table 4) five homogenous groups (Recall, BLEU-1, F-measure), (*BLEU-4*), (*BLEU-3*), (*BLEU-2*) and (*Precision*) were identified in terms of the score of automatic evaluation of MT. Statistically significant differences were proved at the 5 % significance level in the score of automatic evaluation of MT between *BLEU-1*/*BLEU-2*/*BLEU-3* and others as well as between *Precision* and others.

The means that the error plot (Fig. 2) depicts means and confidence intervals of metrics of accuracy. The plot visualizes homogeneous groups as well as differences between examined metrics.

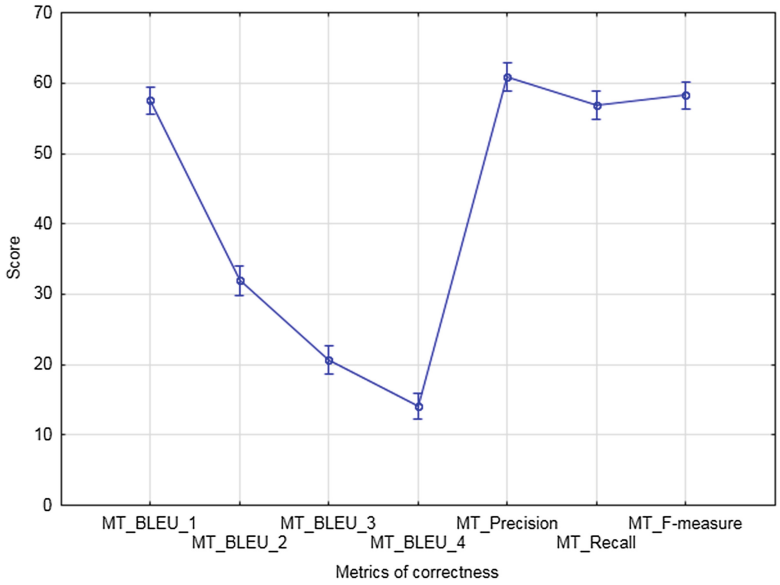

Fig. 2. The means with error plot for intrinsic metrics of accuracy.

Similarly to the metrics of error rate, metrics of accuracy have a significant impact on the quality of MT evaluation, except for metrics *BLEU-1* and *F-measure*. The metrics *BLEU-1* and *F-measure* were identified as redundant metrics of accuracy for MT evaluation.

5 Conclusion and Future Direction

Williams [37] claimed that the techniques and methods for translation quality assessment (TQA) must pass validity and reliability tests if we want TQA procedures to be as objective as possible.

For this reason, we carried out the evaluation of automatic intrinsic metrics for MT evaluation. Evaluation was realized over the textual data obtained from machine translation. Translation was done by MT system (free online machine translation service) and the translation direction was from Slovak (morphologically complex language and low resource) into English.

We showed a way how to identify relevant and redundant automatic metrics for MT evaluation. We presented two approaches to the identification, using the analysis of reliability and using entropy. We used three coefficients of reliability – *Cronbach's alpha*, *Standardized alpha* and *entropy* – to estimate reliability. All estimations were very similar, i.e. individual automatic metrics for MT evaluation have the same variability. The metrics of automatic evaluation have a significant impact on the quality evaluation of MT, except *BLEU-1* and *F-measure* (it was showed that both metrics are redundant in comparison to others).

In future work, we would apply automatic intrinsic metrics for the evaluation of MT output into our translation quality assessment model from Slovak to English and vice versa. In addition, for automatic errors identification and classification of MT output using these metrics which are interconnected to specific morphological and syntactical errors, as well as for the MT evaluation based on POS tagging and for the development of a tool for automatic error detection based on these metrics and morphological annotation of the reference, hypothesis and post-edited MT output.

Above that, MT systems and evaluation of their performance in the context of the inflectional Slovak language has not yet been investigated, which makes the research purposeful and innovative.

Acknowledgments. This work was supported by the Slovak Research and Development Agency under the contract No. APVV-14-0336 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contracts No. VEGA-1/0559/14.

References

1. Liu, C., Dahlmeier, D., Ng, H.T.: Better evaluation metrics lead to better machine translation. In: Proceedings of Conference Empirical Methods in Natural Language Processing, pp. 375–384 (2011)
2. Babych, B., Hartley, A., Sharoff, S.: Translating from under-resourced languages: comparing direct transfer against pivot translation. In: Proceedings of the MT Summit XI. Citeseer, Copenhagen (2007)
3. Adly, N., Al Ansary, S.: Natural Language Processing and Information Systems. Springer, Heidelberg (2010)
4. Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M., Schmidt, P.: Evaluation of a machine translation system for low resource languages: METIS-II (2008)
5. Vilar, D., Xu, J., D'Haro, L.F., Ney, H.: Error analysis of statistical machine translation output. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06), pp. 697–702, Genoa (2006)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL 2002, p. 311. Association for Computational Linguistics, Morristown, NJ, USA (2001)
7. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (ACL 2005), pp. 65–72, Michigan (2005)
8. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, pp. 138–145 (2002)
9. Koehn, P.: Statistical Machine Translation. Cambridge University Press, Cambridge (2010)
10. 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)

11. Munkova, D., Munk, M.: An automatic evaluation of machine translation and Slavic languages. In: 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–5. IEEE (2014)
12. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between European languages (2006)
13. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (Meta-) evaluation of machine translation (2007)
14. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 70–106. ACL (2008)
15. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 workshop on statistical machine translation (2009)
16. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.F.: Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, pp. 17–53. Association for Computational Linguistics (2010)
17. Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O.F.: Findings of the 2011 workshop on statistical machine translation. In: Proceedings of Sixth Workshop Statistical Machine Translation, pp. 22–64 (2011)
18. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation (2012)
19. Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2013 workshop on statistical machine translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation. pp. 1–44. Association for Computational Linguistics, Sofia, Bulgaria (2013)
20. Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A.: Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 12–58. Association for Computational Linguistics, Stroudsburg, PA, USA (2014)
21. Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., Turchi, M.: Findings of the 2015 workshop on statistical machine translation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 1–46. Association for Computational Linguistics, Stroudsburg, PA, USA (2015)
22. 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 2004, p. 605. Association for Computational Linguistics, Morristown, NJ, USA (2004)
23. Han, A.L.-F., Wong, D.F., Chao, L.S., He, L., Lu, Y.: Unsupervised quality estimation model for English to German translation and its application in extensive supervised evaluation. Sci. World J. **2014**, 760301 (2014)
24. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H.: Accelerated DP based search for statistical translation (1997)
25. Munková, D., Munk, M., Adamová, L.: Modelling of language processing dependence on morphological features. In: Trajkovik, V., Anastas, M. (eds.) ICT Innovations 2013, pp. 77–86. Springer International Publishing, Heidelberg (2014)
26. Munková, D., Munk, M., Vozár, M.: Data pre-processing evaluation for text mining: transaction/sequence model. Procedia Comput. Sci. **18**, 1198–1207 (2013)

27. Munková, D., Munk, M., Adamová, L.: Influence of Stop-words removal on sequence patterns identification within comparable corpora. In: Trajkovik, V., Anastas, M. (eds.) ICT Innovations 2013: ICT Innovations and Education. Advances in Intelligent Systems and Computing, pp. 67–76. Springer International Publishing, Heidelberg (2014)
28. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. Proc. RANLP **2005**, 590–596 (2005)
29. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual meeting on Association for Computational Linguistics, pp. 169–176. Association for Computational Linguistics, Morristown, NJ, USA (1991)
30. Tóth, K., Farkas, R., Kocsor, A.: Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm. Acta Cybern. **18**, 463–478 (2008)
31. Melamed, I.D.: Models of translational equivalence among words. Comput. Linguist. **26**, 221–249 (2000)
32. Melamed, I.D.: Statistical machine translation by parsing. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL 2004, p. 653–661. Association for Computational Linguistics, Morristown, NJ, USA (2004)
33. Moore, R.C.: A discriminative framework for bilingual word alignment. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT 2005, pp. 81–88. Association for Computational Linguistics, Morristown, NJ, USA (2005)
34. Yu, X., Wu, J., Zhao, W.: Dictionary-based Chinese-Tibetan sentence alignment. In: 2010 International Conference on Intelligent Computing and Integrated Systems, pp. 489–493. IEEE (2010)
35. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. **5**, 3 (2001)
36. Lima, C.F.L., Assis, F.M., Souza, C.P.: A Comparative study of use of Shannon, Rényi and Tsallis entropy for attribute selecting in network intrusion detection. In: Yin, H., Costa, J.A. F., Barreto, G. (eds.) IDEAL 2012. LNCS, vol. 7435, pp. 492–501. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32639-4_60
37. Williams, M.: Translation quality assessment. Mutatis Mutandis **2**, 3–23 (2009)