Towards the Use of Entropy as a Measure for the Reliability of Automatic MT Evaluation Metrics

Michal Munk[a] , Dasa Munkova[b] and Lubomir Benko[c,*]
[a]*Department of Informatics, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia*
[b]*Department of Translation Studies, Constantine the Philosopher University in Nitra, Stefanikova 67, 949 74 Nitra, Slovakia*
[c]*Institute of System Engineering and Informatics, University of Pardubice, Studentska 95, 532 10 Pardubice, Czech Republic*

Abstract. The study describes an experiment with different estimations of reliability. Reliability reflects the technical quality of the measurement procedure such as an automatic evaluation of Machine Translation (MT). Reliability is an indicator of accuracy, the reliability of measuring, in our case, measuring the accuracy and error rate of MT output based on automatic metrics (precision, recall, f-measure, Bleu-n, WER, PER, and CDER). The experiment showed metrics (Bleu-4 and WER) that reduce the overall reliability of the automatic evaluation of accuracy and error rate using entropy. Based on the results we can say, that the use of entropy for the estimation of reliability brings more accurate results than conventional estimations of reliability (Cronbach's alpha and correlation). MT evaluation, based on n-grams or edit distance, using entropy could offer a new view on lexicon-based metrics in comparison to commonly used ones.

Keywords: Entropy, Machine translation, Reliability estimation, Quality, Automatic MT evaluation

1. Introduction

Information entropy was first introduced by Shannon [5] and used in different fields of informatics. Entropy is a measure of disorder, where lower entropy means an order, and on the other hand a higher entropy disorder. Following Shannon´s definitions [5], entropy can be used as a measure of uncertainty in a data set (a discrete random variable X with values $x_1, \ldots, x_n$). In Machine Translation (MT) evaluation, entropy can reflect the distribution of matched words, i.e. a lower entropy, a more fluent MT output and on the other hand, a higher entropy, a less fluent MT output [12]. Yu et al. [12] denote that entropy can sufficiently reflect the fluency of MT output. MT evaluation plays a crucial role in the development of MT. There are many approaches to MT evaluation, from fully automated quality scoring to manual or human assessment of the quality of MT output. In most evaluation approaches translation quality is viewed as an optimal compromise between adequacy (the degree of meaning preservation) and fluency (correctness of target language) [3]. Approaches to manual or human evaluation of MT, requiring human translator knowledge, assess the quality of MT output along the two axes of target language correctness and semantic fidelity, such as ranking, scales, error analysis, or post-editing [18]. Compared to automatic MT evaluation, which is not only fast and cheap but reusable and language-independent; manual evaluation is regarded as the most reliable but time and labor consuming and not re-usable. Papineni et al. [13] stated that manual evaluation is too slow and time consuming for the development of MT systems, for which

[*] Corresponding author. E-mail: lubomir.benko@gmail.com.

fast feedback on translation quality is extremely important. Several metrics of automatic MT evaluation were developed and applied to a large amount of data. Mostly, they try to estimate the similarity (matches) between MT output assessed (hypothesis) and one or more human translations (references). According to the information type, automatic evaluation can be carried out based on statistical principles, using lexicon-based methods (n-grams or edit distance) such as *BLEU*, *NIST*, *METEOR*, *PER*, *WER*, *TER*; or based on the use of deep linguistic structures, using syntax-based and semantic-based methods (linguistic features- morphological, syntactic or semantic information). Progress in MT relies on the assessment of MT output through effective approaches to the evaluation of translation quality. Reliable evaluation metrics lead to better machine translation.

Measurement quality, in our case the quality of automatic evaluation of MT output, is given by fundamental indicators of measurements such as objectivity, reliability, and validity. The concept of objectivity is trivial since the results are independent on the researcher regarding the distortion of measurement. Validity was verified through the manual evaluation of MT output [7, 8]. The study is focused on the reliability of automatic evaluation, which represents a significant indicator of measurement quality, particularly the indicator of accuracy and reliability of measurement. The study describes the experiment with different estimations of reliability- the quality of automatic evaluation of MT output. It focuses on the usage of entropy when analyzing the reliability of metrics for automatic MT evaluation. It deals with the improvement of reliability analysis of automatic metrics of MT evaluation using entropy. The improvement is in the implementation of entropy in reliability estimation of automatic MT evaluation. MT evaluation, based on n-grams or edit distance, using entropy could offer a new view on lexicon-based metrics in comparison to commonly used ones.

The rest of the paper is structured as follows: section 2 is related work, section 3 describes entropy and section 4 reliability associated with the traditional methods; section 5 presents experiment where entropy was an alternative estimation for the reliability analysis of MT evaluation. Subsequently, the conclusions and future work are offered in the last section.

2. Related Work

Entropy is a universal measure, which can be applied to any field such as linguistics or natural language processing. Montemurro et al. [19] used a relative entropy to measure the degree of ordering in word sequences from languages belonging to five linguistic families and found out that entropy is an almost constant value for all those families. Yu et al. [12] used entropy as an indicator of fluency for the automatic metrics of MT evaluation. They combined entropy with *BLEU* and *METEOR* metrics and found out, that this combination can improve the performance of these metrics, i.e. effectively measure the fluency of a sentence of MT output. Carl and Schaeffer [16] used word translation entropy to show its effectiveness for expected MT quality of literal translations.

Eetemadi et al. [25] offered a complex survey of data selection methods in machine translation. They also describe works focusing on cross-entropy which has become the most commonly used approach in data selection. Shah et al. [14] used cross-entropy as a feature for quality estimation of a neural MT model which led to large improvements in prediction. Tomeh et al. [20] introduced a novel framework based on maximum entropy for word alignment. Based on the experiment, authors improved the alignment quality and translation quality as measured by standard metrics.

Entropy in the context of the analysis of the reliability of automatic MT evaluation was first used by Munk et al. [17]. The authors dealt with the verification of reliability of automatic MT evaluation. They tried to identify the redundant metrics of automatic MT evaluation. They restricted only to the interpretation of estimates of reliability. The estimates were not mutually compared and assessed, which estimate is the most suitable for assessing the reliability of automatic MT evaluation. Based on the results [17] we examine the use of entropy as an estimation of the reliability of automatic MT evaluation in comparison to traditional/conventional estimates.

3. Entropy

Entropy was first introduced in thermodynamics [22], and it was used to provide a statement of the second law of thermodynamics on the irreversibility of evolution. It was understood that an isolated system could not pass from a state of a higher entropy to a state of a lower entropy [1]. Shannon first introduced entropy as a measure of uncertainty in a discrete distribution in information theory [1]. Mostly, entropy in information theory is defined as a degree of the system's disorder or randomness. Based on Shannon's definition [5, 6, 17], 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 [5, 6, 17] as

$$H(C) = -\sum_{i=1}^{k} p_i \log p_i,$$

the function decreases from infinity to zero and p_i takes values in the range [0,1]. Entropy as a modeling tool was formulated by Jaynes [10] and is known as Maximum entropy.

Entropy can be used in many fields and offers an alternative way to already existing methods. Many authors use entropy measures for fuzzy sets. Farhadinia [2] introduced two general ways to generate entropy measures for linguistic terms. The novel approach is demonstrated by determining objective weights of attributes to solve problems in linguistic term sets. Shen et al. [21] introduced two types of distributed semi-supervised metric learning frameworks which were applied on a SERAPH method. SERAPH is a centralized semi-supervised information-theoretic metric learning algorithm combining the generalized maximum entropy principle with entropy regularization in objective function [21]. Integration of correntropy criterion to the minimum error entropy to form centered error entropy criterion is introduced by Cheng [15]. It compares to the mean square error criterion as the performance index for filter design in the description of high order statistics of error PDF for multipath estimation in non-Gaussian noise [15].

4. Reliability

Reliability is determined by the Cronbach's alpha or correlation. If the coefficient of reliability equals one, the result was not affected by the measurement error. The coefficient of reliability cannot achieve such value using the measuring procedure. Some error occurs in every measurement. Our aim is to reduce this error to a minimum. Reliability reflects the technical quality of the measurement procedure.

For the need of reliability analysis direct and indirect estimates of reliability are commonly used to assess the reliability of the measurement procedure. Indirect estimation of reliability is usually done by correlation coefficient or its modification. Direct estimation of reliability is Cronbach's alpha

$$\hat{\alpha} = \frac{m}{m-1}\left(1 - \frac{\sum s_j^2}{s^2}\right),$$

where m is the number of metrics of measurement procedure, s^2 is the variance of scores of measuring procedure- automatic evaluation of accuracy (error rate), s_j^2 is the variance of the j^{th} metrics of automatic evaluation of accuracy (error rate) of MT output [23].

If the assessments are only incidental and do not reflect the actual quality of machine translation, then the errors are random and uncorrelated. In this case, the sum of the variances equals the variance of the sum, and Cronbach's coefficient alpha equals zero. If all the metrics are reliable and reflect the actual quality of the machine translation, then coefficient alpha equals one. The higher Cronbach's alpha, the more reliable the measuring procedure, in our case the automatic evaluation of accuracy (error rate) of the MT output [24].

We used the following conventional estimations of reliability and entropy to identify the automatic metrics reducing the overall reliability of the automatic MT evaluation:

- Cronbach's alpha coefficient after removing the relevant metric (*Alpha if deleted*), where if the value of the coefficient after removing the relevant metric is higher than the overall reliability of automatic evaluation, then the relevant metric reduces the overall reliability of the automatic evaluation- metrics with a higher reliability when removed, can be considered suspicious;
- the correlation coefficient between the relevant metric and overall score of automatic evaluation (*Metrics-total correlation*), where metrics with a lower level of dependency can be seen as suspicious;
- the coefficient of entropy for the relevant metric (*Metrics-total entropy*), where metrics with a lower level of entropy can be regarded as suspicious.

5. Experiment

The objective of the experiment is an identification of metrics that reduce the overall reliability of the automatic evaluation of accuracy and error rate of MT output using entropy, Cronbach's alpha, and correlation. Reliability is an indicator of accuracy, the reliability of measuring, in our case for the measuring of accuracy and error rate of MT output based on automatic evaluation metrics.

5.1. Dataset

We created a dataset covering two translation directions- one direction, a translation from Slovak (SK) into English (EN), and the second direction, a translation from Slovak (SK) into German (DE). Slovak belongs to a highly inflected languages contrary to English, which belongs to analytical languages, i.e. one English noun has six forms in Slovak differing only in suffixes, German has four cases and Slovak six, or different word order, English/German has a strict word order (SVO) in comparison to Slovak's loose word order, or in Slovak the agent does not have to be expressed compared to English or German). The examined text was original, written in Slovak, consisting of 360 sentences and translated to English and German by a statistical machine translation system, Google Translate. The examined dataset is limited to 360 sentences because it was obtained from one-day workshop focusing on post-editing and manual evaluation of MT output (only 360 sentences were post-edited and manually assessed during one day). We chose these directions to obtain higher scores of automatic metric *BLEU*. *BLEU* metric as a measure for translation quality assessment is not suitable for translation into inflectional languages.

5.2. Metrics

In this study, similar to Munk et al. experiment [17], we will focus only on automatic metrics based on lexicon methods.

Precision and *Recall* are based on the closeness of the hypothesis (MT output) with the reference (human translation), similar to bag-of-words, i.e. regardless of the position of the word in a sentence.

Precision (*P*) is a measure of how many correct words are present in the hypothesis in regard of reference

$$P = \frac{number\ of\ correct\ words\ in\ hypothesis}{number\ of\ words\ in\ hypothesis}.$$

It is a proportion of words in MT output that are present in the reference translation.

Recall (*R*) is the number of correct words in MT output divided by the number of words of reference, i.e. proportion of all words in reference that are correct

$$R = \frac{number\ of\ correct\ words\ in\ hypothesis}{number\ of\ words\ in\ \ reference}.$$

It is a proportion of words in the reference that are present in the MT output.

F-measure (*F₁*) is a harmonic mean of *precision* and *recall*

$$F_1 = \frac{2PR}{P+R}.$$

BLEU [13] is a geometric mean of n-gram *precisions* and the second part is a *brevity penalty* (*BP*), i.e. length-based penalty to prevent very short sentences as compensation for inappropriate translation.

$BLEU\ (n) = exp \sum_{n=1}^{N} w_n \log p_n \times BP$, where w_n is weights for different p_n.

$BP = \begin{cases} 1, & if\ r > r \\ e^{1-\frac{r}{h}}, & if\ h \leq r \end{cases}$, where *r* is a reference of a hypothesis *h*.

BLEU represents two features of translation quality- *adequacy* and *fluency* by calculating words or lexical *precision*.

Other widely used evaluation metrics are based on edit distance such as *PER*, *WER*, or *CDER*.

WER [26] is based on the edit distance (edit operations) and does not allow reordering of words. It accounts the Levenshtein distance between a hypothesis and a reference. The minimum number of edit operations (insertions, substitutions, and deletions of words necessary to transform the hypothesis into the reference) is divided by the number of words in the reference.

$WER(h, r) = \frac{min_{e \in E(h,r)}(I+D+S)}{|r|}$, where *r* is a reference of a hypothesis h, I - insert(e), D - delete(e), S - substitute(e), and $min_{e \in E(h,r)}$ is a minimal sequence of transformed word (insertions, substitutions and deletions).

PER [4] is based on *WER* but ignores the ordering of the words in a sentence, i.e. word order is not taken into account.

CDER [11,17] is a measure oriented towards *recall*, but 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 requires both, hypothesis and reference to be covered completely and disjointly. Only words in the reference must be covered only once, while in the hypothesis they can be covered zero, one or more times.

PER, *WER*, and *CDER* metrics are called metrics of error rate, i.e. the higher the values of these metrics, the lower the translation quality. On the other hand, metrics *Precision*, *Recall*, *F-measure*, and *BLEU-n* are called metrics of accuracy, i.e. the higher the values of these metrics, the better the translation quality.

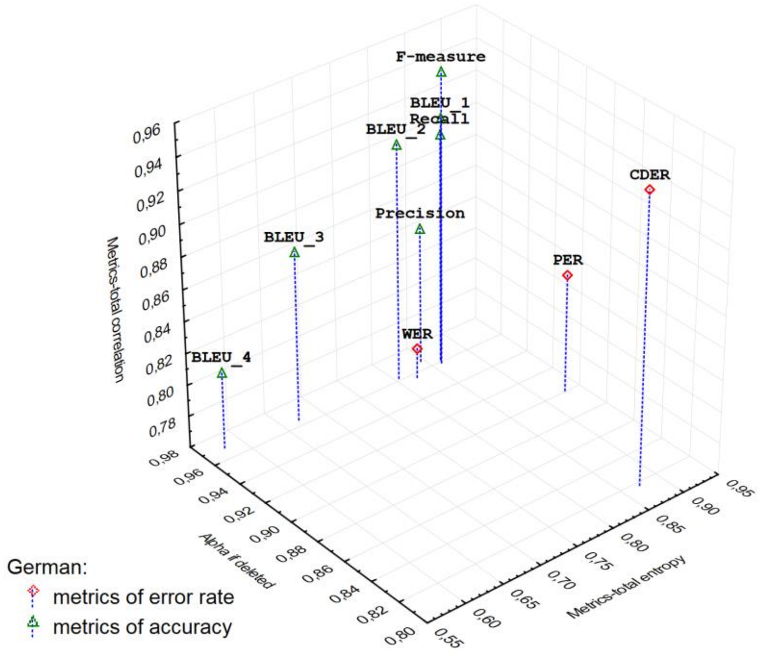

Fig. 1. Categorized 3D scatterplot of reliability estimations of examined metrics of German MT.

Automatic sentence alignment was carried out using the algorithm and software hunalign, based on dictionary and length based methods, which both successfully amalgamates [9].

We used our tools to compute, based on the comparison of the hypothesis with reference, the automatic scores for the metrics of error rate (*PER*, *WER*, and *CDER*) and metrics of accuracy (*Precision, Recall, F-measure,* and *BLEU-n*).

5.3. Results

Entropy was calculated for each sentence. It was calculated from the value of the given metric of error rate added to the corresponding value of the metric of accuracy. The sum of the metrics should always be 100 %. The total entropy was calculated as the average of the entropy of all sentences, separately for metrics of error rate and metrics of accuracy. If the entropy is closer to 1, then the system tends to disorder. The results of the entropy relate with the coefficient of reliability estimate.

A similar method was applied to the calculation of entropy estimations for metrics of accuracy (Figure 1, Figure 2) as in the case of metrics of error rate. The average entropy was also used, and the results related to the reliability analysis with negligible variations.

The categorized 3D Scatterplot (Figure 1, Figure 2) visualizes distance between metrics of accuracy and error rates based on reliability estimations: *Metrics-total entropy*, *Metrics-total correlation,* and *Alpha if deleted*. The correlations between the estimates and the sums score (without the respective metric) are shown on the z-axis (*Metrics-total correlation*). The y-axis shows the resultant *Cronbach's alpha* value if the respective metrics were to be removed (*Alpha if deleted*) and the x-axis shows the calculated entropy for each specific metric (*Metrics-total entropy*).

In the case of metrics of accuracy (Figure 1, Figure 2), the most deviating from others is *BLEU_4* metric, which achieved the lowest entropy and correlation, and the highest reliability after removing the respective metric. In the case of metrics of error rate (Figure 1, Figure 2), the most deviating is *WER* metric, which achieved the lowest entropy and correlation, and the highest reliability after removing the respective metric.

In the case of machine translation into English (Figure 2) metrics *Precision* and *BLEU_3* deviate the most when we do not take into account *BLEU_4*. These metrics are characterized with a lower correlation with a

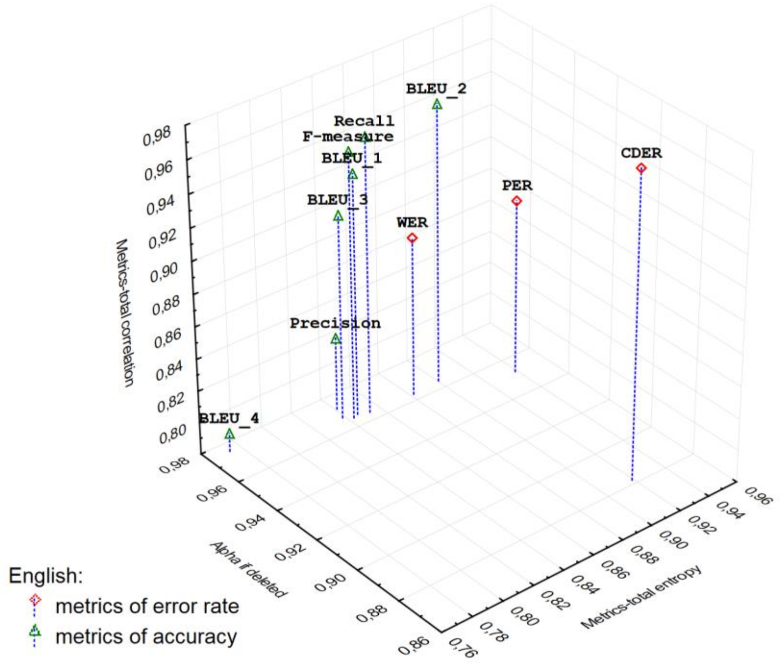

Fig. 2. Categorized 3D scatterplot of reliability estimations of examined metrics of English MT.

total score and with a high reliability after their removing. However, they achieve about the same entropy as the rest of the metrics of accuracy, which can be considered to be reliable.

In the case of machine translation into German (Figure 1), these deviations are more visible. Similarly, *Precision* and *BLEU_3* are characterized with a low correlation with a total score and with a high reliability after their removing. However, *Precision* achieves about the same entropy as the rest of the metrics of accuracy, which can be considered to be reliable based on *Alpha if deleted* and *Metrics-total correlation*.

In the case of metrics of accuracy it was identified a very large directly proportional dependency (r_{GERMAN} = 0.766; $r_{ENGLISH}$ = 0.776) between entropy (*Metrics-total entropy*) and the correlation (*Metrics-total correlation*), i.e. values are changing together in the same direction. Similarly, between entropy (*Metrics-total entropy*) and the resultant Cronbach's alpha value, if the particular metric was to be removed (*Alpha if deleted*), it was identified a very large degree of dependency but inversely proportional (r_{GERMAN} = -0.732; $r_{ENGLISH}$ = -0.804), i.e. values are changing together in the opposite direction.

In the case of machine translation to German, there was identified a high degree of dependency (*Multiple R* = 0.928) between entropy and conventional estimations of reliability (*Metrics-total correlation*, *Alpha if deleted*). Using the conventional estimations of reliability, we can explain the variability of entropy to 86% (*Squared multiple R* = 0.862). Multiple correlation coefficient is statistically significant at the 0.05 significant level (*F*(2,4) = 12.484; *p* = 0.0191). Similarly, in the case of metrics of the accuracy of machine translation to English, it was identified a high degree of dependency (*Multiple R* = 0.940; *F*(2, 4) = 15.147; *p* = 0.0136) between entropy and conventional estimations of reliability. In this case, the conventional estimate explains the variability of entropy up to 88% (*Squared multiple R* = 0.883).

The following graphs visualize differences in measures of reliability among examined MT metrics of accuracy. If we look deeply at the results for German MT output (Figure 3), we can see based on the *Alpha if deleted* estimate, *BLEU_4* is identified as metrics decreasing the total score of the reliability of automatic evaluation of the accuracy of MT (*Cronbach's alpha* = 0.9681). Similarly, based on *Metrics-total correlation* and *Metrics-total entropy* estimates (Figure 4, Figure 5), we identified *BLEU_4* as suspicious. However, in the case of *Alpha if deleted* estimate (Figure 3) another highest score was achieved in *Precision*

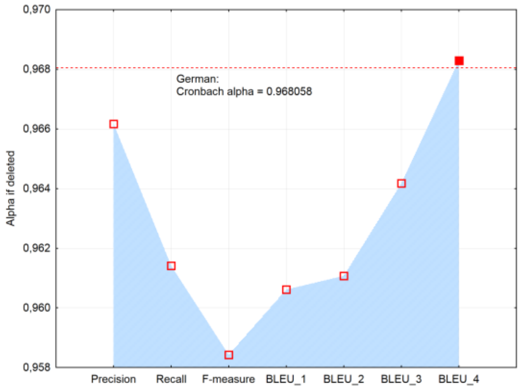

Fig. 3. Line plot of Alpha if deleted of examined metrics of German MT.

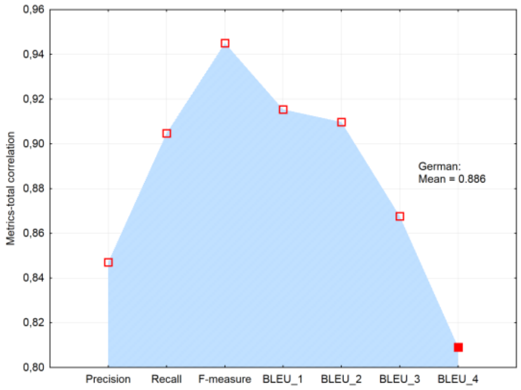

Fig. 4. Line plot of Metrics-total correlation of examined metrics of German MT.

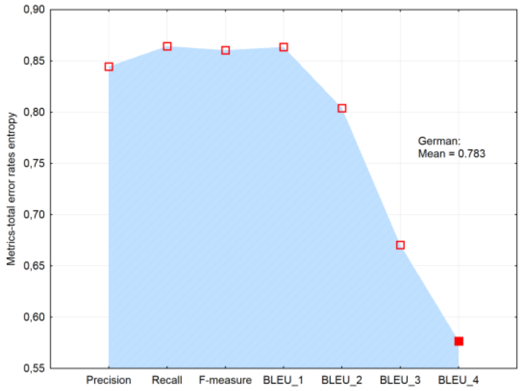

Fig. 5. Line plot of Metrics-total error rates entropy of examined metrics of German MT.

and afterward *BLEU_3*. Also, in the case of *Metrics-total correlation* estimate (Figure 4), the lowest score was achieved in *Precision* and then in *BLEU_3*. However, in the case of *Metrics-total entropy* estimate (Figure 5) it was in reverse order, i.e. the next lowest score was achieved in *BLEU_3*, *BLEU_2*, and then in *Precision*.

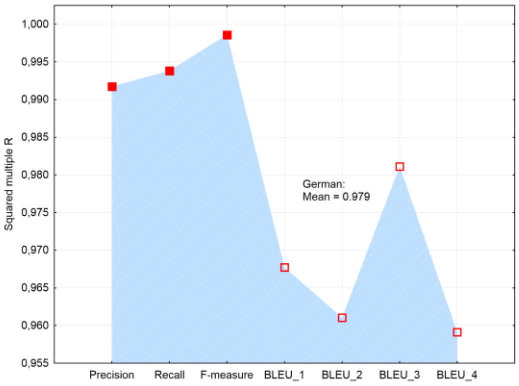

Fig. 6. Line plot of Squared multiple R of examined metrics of German MT.

In the case of English machine translation, the results are similar. Based on *Alpha if deleted*, *Metrics-total correlation* and *Metrics-total entropy* estimates (Figure 7, Figure 8, Figure 9) *BLEU_4* was identified as suspicious- decreasing the overall reliability of automatic evaluation of the accuracy of MT output. After *Precision* removing, the total score of the reliability of automatic evaluation of the accuracy of MT output has not changed (Figure 7). Additionally, in the case of *Metrics-total correlation* estimate (Figure 8), *Precision* together with the *BLEU_4* have achieved the lowest scores. From this point of view, *Precision* seems to be suspicious. However, in the case of *Metrics-total entropy* estimate (Figure 9), *Precision* has achieved almost the same score as metrics, which are regarded as reliable based on *Alpha if deleted* and *Metrics-total correlation* estimates.

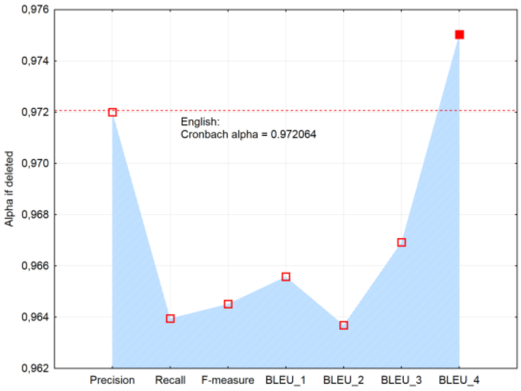

Fig. 7. Line plot of Alpha if deleted of examined metrics of English MT.

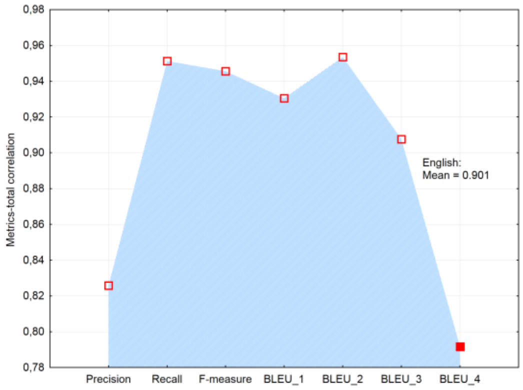

Fig. 8. Line plot of Metrics-total correlation of examined metrics of English MT.

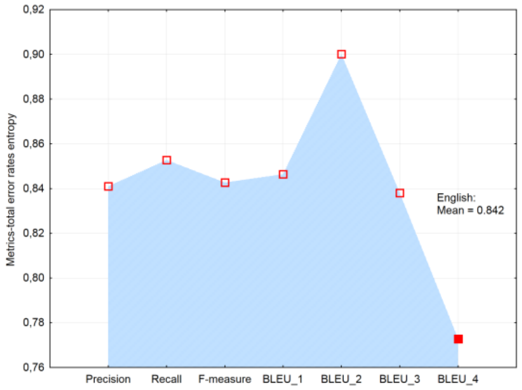

Fig. 9. Line plot of Metrics-total error rates entropy of examined metrics of English MT.

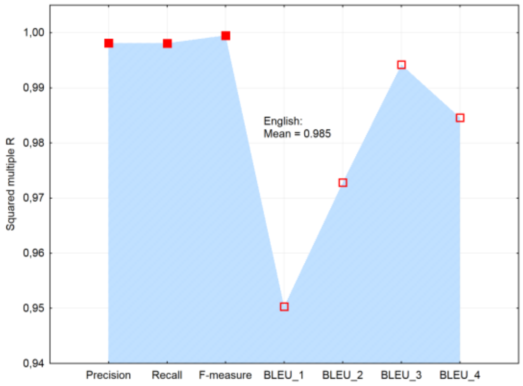

Fig. 10. Line plot of Squared multiple R of examined metrics of English MT.

Metrics-total entropy estimate is more sensible for the internal consistency of measurement procedure of automatic evaluation of the accuracy of MT output given that, the metric *F-measure* is the harmonic mean of *Precision* and *Recall*. In other words, the variability of *Precision/Recall/F-measure* (Figure 6, Figure 10) is explained to more than 99% by the rest of metrics (Squared multiple R > 0.99).

Similar results were also achieved for the metrics of error rates, where it was achieved a moderate degree of dependency between the entropy and conventional estimations of reliability.

The analysis of reliability of automatic metrics characterizing the error rate of MT evaluation showed that the *WER* metric deviates the most from the other automatic metrics of automatic MT evaluation (*PER* or *CDER*) in translation quality assessment. It is understandable seeing that, the *WER* metric is very strict to syntax errors (word order).

In the case of automatic metrics of accuracy, it was also showed that *BLEU_4* metric deviates the most from the other metrics. We explain this by the fact that *BLEU_4* metric measures a score of a sequence of four words (including articles and prepositions) and sometimes it is very complicated to achieve an output with such sequence by MT systems.

6. Conclusions

The aim of this experiment was to assess the reliability of the automatic evaluation of machine translation for inflectional languages using traditional estimates and entropy.

This paper introduced entropy as an alternative means of reliability estimation of machine translation evaluation. In this case, entropy was implemented into the system, where as an estimate of reliability for automatic evaluation of the machine translation was used.

Three different estimations of reliability were used – Cronbach’s alpha, correlation, and entropy – to estimate reliability. In the experiment, entropy was proved to be an equivalent (appropriate) alternative to the former ones.

Besides, entropy has been shown to be more sensitive to the internal consistency of automatic evaluation of accuracy of MT output and has accurately identified the metrics suspicion. It is common that the automatic metrics of accuracy and error rates are implemented in the systems of MT evaluation. It allows using entropy for the reliability estimation of automatic MT evaluation. Based on the results we can say, that the use of entropy for the estimation of reliability brings more accurate results than conventional estimations of reliability. Entropy has also correctly identified suspicious metrics (*Bleu-4* and *WER* metrics deviating the most from the total measurement of accuracy/error rate of MT output) like conventional estimations of reliability. However, in the case of remaining metrics that deviated less from the measurement, entropy produced more accurate results.

In any case, we recommend combining several methods of reliability estimation. If the results are the same, we can consider them as robust.

However, in the case of larger textual sets, the calculation of entropy for every sentence can become very tedious. The future work in this area could be focused on improving the calculation speed in the case of a large dataset. Despite that, we were able to evaluate the reliability of metrics using entropy, and the results relate to the results of the reliability analysis.

We believe that the presented research will be an inspiration for the use of entropy in various fields and stages of data processing.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-14-0336 and Cultural and Educational Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the contracts No. KEGA-041UK-4/2017.

References

[1] A. Holzinger, M. Hörtenhuber, C. Mayer, M. Bachler, S. Wassertheurer, A.J. Pinho, and D. Koslicki, On Entropy-Based Data Mining, in: A. Holzinger, and I. Jurisica (Eds.), Interact. Knowl. Discov. Data Min. Biomed. Informatics State-of-the-Art Futur. Challenges, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014: pp. 209–226. doi:10.1007/978-3-662-43968-5_12.

[2] B. Farhadinia, Determination of entropy measures for the ordinal scale-based linguistic models, Inf. Sci. (Ny). 369 (2016) 63–79. doi:10.1016/j.ins.2016.06.002.

[3] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, (Meta-) evaluation of machine translation, Proc. Second Work. Stat. Mach. Transl. (2007) 136–158. http://dl.acm.org/citation.cfm?id=1626373 (accessed August 8, 2017).

[4] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, Accelerated DP based search for statistical translation., in: Eur. Conf. Speech Commun. Technol., Rhodes, Greece, 1997: pp. 2667–2670.

[5] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mob. Comput. Commun. Rev. 5 (2001) 3. doi:10.1145/584091.584093.

[6] C.F.L. Lima, F.M. de Assis, and C.P. de Souza, A Comparative Study of Use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection, in: Springer Berlin Heidelberg, 2012: pp. 492–501. doi:10.1007/978-3-642-32639-4_60.

[7] D. Munkova, and M. Munk, An automatic evaluation of machine translation and Slavic languages, in: 2014 IEEE 8th Int. Conf. Appl. Inf. Commun. Technol., IEEE, 2014: pp. 1–5. doi:10.1109/ICAICT.2014.7035992.

[8] D. Munková, and M. Munk, Automatic Metrics for Machine Translation Evaluation and Minority Languages, in: Mediterr. Conf. Inf. Commun. Technol. MedCT 2015, Springer, Cham, Saidia, 2016: pp. 631–636. doi:10.1007/978-3-319-30298-0_69.

[9] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, Parallel corpora for medium density languages, Proc. RANLP 2005. (2005) 590–596.

[10] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620.

[11] G. Leusch, N. Ueffing, and H. Ney, CDER: Efficient MT Evaluation Using Block Movements, in: Proc. 11th Conf. Eur. Chapter Assoc. Comput. Linguist. (EACL 2006), 2006.

[12] H. Yu, X. Wu, W. Jiang, Q. Liu, and S. Lin, Improve the Evaluation of Fluency Using Entropy for Machine Translation Evaluation Metrics, (2015). http://arxiv.org/abs/1508.02225 (accessed August 8, 2017).

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proc. 40th Annu. Meet. Assoc. Comput. Linguist., Philadelphia, 2002: pp. 311–318.

[14] K. Shah, F. Bougares, L. Barrault, and L. Specia, SHEF-LIUM-NN: Sentence level Quality Estimation with Neural Network Features, in: Proc. First Conf. Mach. Transl. Vol. 2, Shar. Task Pap., Association for Computational Linguistics, Berlin, 2016: pp. 838–842.

[15] L. Cheng, M.F. Ren, and G. Xie, Multipath Estimation Based on Centered Error Entropy Criterion for Non-Gaussian Noise, IEEE Access. 4 (2016) 9978–9986. doi:10.1109/ACCESS.2016.2639049.

[16] M. Carl, and M. Schaeffer, Word Transition Entropy as an Indicator for Expected Machine Translation Quality, in: Proc. Work. Autom. Man. Metrics Oper. Transl. Eval. MTE 2014, ELRA, Paris, 2014: pp. 45–50.

[17] M. Munk, D. Munková, and Ľ. Benko, Identification of Relevant and Redundant Automatic Metrics for MT Evaluation, in: Multi-Disciplinary Trends Artif. Intell. (MIWAI 2016) B. Ser. Lect. Notes Comput. Sci., Springer International Publishing, Cham, 2016: pp. 141–152. doi:10.1007/978-3-319-49397-8_12.

[18] M. Vela, A.-K. Schumann, and A. Wurm, Human Translation Evaluation and its Coverage by Automatic Scores, in: Proc. Work. Autom. Man. Metrics Oper. Transl. Eval. MTE 2014, ELRA, Paris, 2014: pp. 20–30.

[19] M. A. Montemurro, D.H. Zanette, S. Havlin, M. Simons, and H. Stanley, Universal Entropy of Word Ordering Across Linguistic Families, PLoS One. 6 (2011) e19875. doi:10.1371/journal.pone.0019875.

[20] N. Tomeh, A. Allauzen, and F. Yvon, Maximum-entropy word alignment and posterior-based phrase extraction for machine translation, Mach. Transl. 28 (2014) 19–56. doi:10.1007/s10590-013-9146-4.

[21] P. Shen, X. Du, and C. Li, Distributed Semi-Supervised Metric Learning, IEEE Access. 4 (2016) 8558–8571. doi:10.1109/ACCESS.2016.2632158.

[22] R. Clausius, On the Motive Power of Heat, and on the Laws which Can be Deduced from it for the Theory of Heat, Dover, 1960. https://books.google.sk/books?id=WEzmcQAACAAJ.

[23] R.F. DeVellis, Scale Development, Sage Publ. (1991) 24–33.

[24] R.F. DeVellis, Scale Development: Theory and applications, Sage, Los Angeles, 2012.

[25] S. Eetemadi, W. Lewis, K. Toutanova, and H. Radha, Survey of data-selection methods in statistical machine translation, Mach. Transl. 29 (2015) 189–223. doi:10.1007/s10590-015-9176-1.

[26] S. Nießen, F.J. Och, G. Leusch, and H. Ney, An evaluation tool for machine translation: Fast evaluation for MT research, in: Proc. 2nd Int. Conf. Lang. Resour. Eval., 2000: pp. 39–45.