Using binary classification to evaluate the quality of machine translators

Ran Li[1,*], Yihao Yang[1], Kelin Shen[2], Mohammad Hijji[3]

[1]*School of Computer and Information Technology, Xinyang Normal University, Xinyang, China*
[2] *School of Foreign Languages, Xinyang Agriculture and Forestry University, Xinyang, China*
[3] *Industrial Innovation and Robotic Center (IIRC), University of Tabuk, Tabuk 47711, Saudi Arabia*
**Corresponding author: liran@xynu.edu.cn*

Abstract

Machine translators have become increasingly popular and currently play an important role because of their great assistance in cross-cultural communication. However, machine translators often produces some unnatural texts, and an evaluation of machine translators is thus needed to avoid the abuse of machine-translated texts. This paper presents the use of binary classification to evaluate the quality of machine translators without references. First, we construct a large-scale dataset including human-generated texts and machine-translated texts. Second, the dataset is used to train the multiple binary classifiers, e.g., decision tree, random forest, extreme gradient boosting, support vector machines, logistic regression, etc. Finally, these trained classifiers constitute the ensemble model by majority voting, and this ensemble model is used to evaluate the qualities of machine-translated texts. Experimental results show that the proposed evaluation method better measures the qualities of translated texts by some commercial machine translators.

Keywords: Binary classification, ensemble model, machine translator, majority voting, quality evaluation.

1. Introduction

Because machine translation research is thriving, machine translators are used to translate the native language into the target language (Matthew *et al.*, 2006). However, different machine translation algorithms can provide a wide diversity of target translations of a single source sentence. If the quality of the machine translator is terrible, the translation can be misleading, producing an inaccurate result. This necessitates that the machine translation output be of human translation quality. Therefore, to choose a suitable translator, many scholars are working to develop trustworthy ways to evaluate machine translators.

Previous methods for machine translator evaluation are mostly based on comparison reference translation, e.g., METEOR (Banerjee & Lavie, 2005; Chung, 2020), BLEU (Ehud, 2018; Papineni *et al.*, 2002), and TER (Matthew, 2010). To score machine translation output, these traditional methods use the sentence-by-sentence procedure. Joty *et al.* (2014) and Mann *et al.* (1988) proposed a method to assess the accuracy of machine translation by using a discourse tree and defined two effective similarity indices. Alexandra & Alexey (2011) designed a phrase-based detection method to filter machine translation documents from network data. The evaluation strategies (Keiji & Sugaya, 2001; Spencer *et al.*, 2011) made a comparison between the parallel text's reference translation and the translator output. The National Institute of Standards and Technology (NIST) method (Yao *et al.*, 2006) is proposed to compare machine translation output with a reference translation in terms of statistical word frequency. Iida & Tokunaga (2012) proposed a method to use decision models to assess the quality of text translation. Kexin (2020) provided a set of automatic machine translation evaluation metrics, which are evaluated through one-to-many alignment. Another sort of quality evaluation takes into account the characteristics of common

and discourse relationships between machine and human translations. Yasuhiro *et al.* (2001) employed different edit distances to analyse machine translation outputs by reference translation automatically. As a similarity metric, Turian *et al.* (2003) employed the maximum mapping size of the translator output and reference translation. Lin & Kan (2011) and Wong & Kit (2011) proposed an evaluation index method for machine translation based on the cohesiveness of words. Comelles *et al.* (2010) and Hardmei & Federico (2010) proposed a machine translation evaluation method based on textual meaning that considers the characteristics of semantic relationships and relationships between discourses to evaluate the quality of machine translation. Karamanis *et al.* (2004) and Grosz *et al.* (1995) proposed a coherence machine translation evaluation method using text centre transition. Some automatic methods to evaluate the quality of machine translation have aroused widespread concern among researchers. Ayala & Chen (2017) focused on supervised neural network models to better understand and anticipate the accuracy and fluency of English-Spanish machine translation. Five tasks were conducted using three classifiers in Weka, an open-source machine learning tool. Based on the 'Skopos theory', Cai & Zhou (2016) analysed the quality of translation of translators. Munkova (2018) detected errors in machine translation using residuals and metrics of automatic evaluation. Khan *et al.* (2016) presented a brief introduction of basic types of linguistic knowledge to discuss different existing machine learning models and their classification into different categories. Khan *et al.* (2016) used pattern and semantic analyses to improve the existing unsupervised opinion target extraction technology. Recently, the use of reference translation to evaluate machine translation quality has attracted much attention (Barzilay & Lapata, 2005). Cindy & Martin (2020) described a machine translation dataset for evaluating machine translation between any of the official languages. Shimanaka *et al.* (2019) used BERT regression to evaluate machine translation. Anurag *et al.* (2020) proposed various metrics for evaluating machine translation based on statistical analysis.

Although the existing methods provide credible evaluation results, they are unable to assess the quality of machine translators in the absence of a reference translation. To address this defect, we propose using binary classification to evaluate the machine translator. The main idea is to treat the evaluation of the machine translator as an issue of binary classification, and the classification accuracy is used as a metric to measure the quality of the machine translator (Wojciech *et al.*, 2021). First, several binary classifiers are used to classify machine-translated and human-generated texts. Second, the majority voting ensemble (Young & Arun, 2021) is used to further improve the classification accuracy. Finally, this ensemble model is used to design the quality evaluation index. The following are the contributions of our work:

- Evaluation method without reference translation. Unlike previous works that use parallel reference translation, our method only requires the monolingual machine translation output as input in the quality assessment stage.
- Majority voting ensemble. The variance is reduced through the ensemble of multiple models to improve the robustness.
- Quality evaluation index. We propose a method to calculate the index of the machine translator, which can compare the performance of various machine translators more intuitively.

The remainder of this work is arranged as follows. Section 2 briefly introduces machine translators and the five binary classifiers used in this evaluation. Section 3 presents the quality evaluation of the machine translator based on the binary classification architecture and working model. Experimental results are provided in Section 4, and Section 5 concludes this paper.

2. Background

2.1 Machine Translators

Machine translation is a versatile and inventive technology that can accomplish a variety of tasks. In several contexts, translation technology has significantly advanced in recent years. Machine translation is the process by which a computer system recognizes the structural aspects and grammar of a source language and automatically generates and converts the text to the destination language. However, because this discernment is based on the automatic recognition of the machine, machine translation has a low quality rate (Sen & Raghunathan, 2018). The goal of machine translation is to examine the vocabulary, syntax, and structure of the origin language and then invoke the system language database to restructure and merge the origin language into a target language that is structurally comparable to the origin language (Duan *et al.*, 2020). It is basically a literal translation process and often shows a lack of knowledge of the origin language. Therefore, the translated texts are rather strict, and grammatical errors may occur. In this paper, the proposed method can be used to evaluate the quality of machine translators for different languages by training binary classifiers using the training dataset generated by the corresponding machine translators. In the experimental stage, we use the Baidu, Bing, Google and Youdao translators as the experimental test subjects for comparing the qualities of the four machine translators, providing help for better translation.

2.2 Binary Classification

In this method, the main idea is to treat the evaluation of the machine translator as an issue of binary classification. We employ a variety of common classifiers, e.g., decision tree (DT) (Coppersmith *et al.*, 1999), random forest (RF) (GUO *et al.*, 2020), extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016; Ogunleye & Wang, 2018; Zhang *et al.*, 2018), support vector machine (SVM) (Rawat *et al.*, 2018; Wang *et al.*, 2018; Zhang *et al.*, 2008), and logistic regression (LR) (Reed & Wu, 2013). These classifiers have low computational complexity in binary classification and provide good accuracy, and they are used in the experiment section to verify the efficacy of our strategy. These common classifiers are briefly described in the following.

- Decision Tree. DT is a classifier model whose child nodes represent results, while leaf nodes represent categories. The training dataset is utilized to create the decision tree, and the best decision tree is obtained. The best decision tree is then used to make decisions for each node in the tree.
- Random Forest. RF is a meta-classifier that can learn and train data to generate many decision trees and make predictions by voting on those trees, which is simple, has low computational complexity, and performs well in classification tasks.
- Extreme gradient boosting. XGBoost is a decision tree-based integrated machine learning method that is based on the gradient boosting decision tree (GBDT), and it also employs the forward step and addition model method to achieve learning optimization.
- Support vector machine. SVM is used to solve binary classification problems. The fundamental model is a linear classifier defined by the maximum interval of the feature space. Interval maximization is the learning method of SVM, which can be characterized as a question of handling convex quadratic programming and is analogous to the regularized hinge loss function minimization issue.
- Logistic regression. LR is a model for generalized linear regression analysis that is often used to estimate the likelihood of something. The essence of LR is to assume that the data follow this distribution and then utilize maximum likelihood estimation to determine the parameters.

3. Proposed Evaluation Method

3.1 Framework Overview

Figure 1 presents the framework of the devised classifier-based machine translator evaluation. In the first module, Chinese-English bilingual abstracts are collected from the bilingual corpus. The Chinese text is translated into English text using four kinds of machine translators, e.g., the Bing, Google, Baidu, and Youdao translators. The four kinds of machine-translated English texts constitute the corresponding dataset of the translator with human-translated English text. The dataset is separated into test and training datasets. The training dataset is used to train the classifier, while the test dataset is used to assess the quality of the machine translator. In the second module, several classifiers, e.g., XGBoost, decision tree, SVM, random forest, and logistic regression, are trained by the text of the training dataset represented by feature vectors. In the third module, we use the test dataset to evaluate the qualities of the four translators. Finally, we obtain the quality evaluation index of four kinds of translators through the above three modules. In the following section, we describe the training classifiers and the quality evaluation in detail.

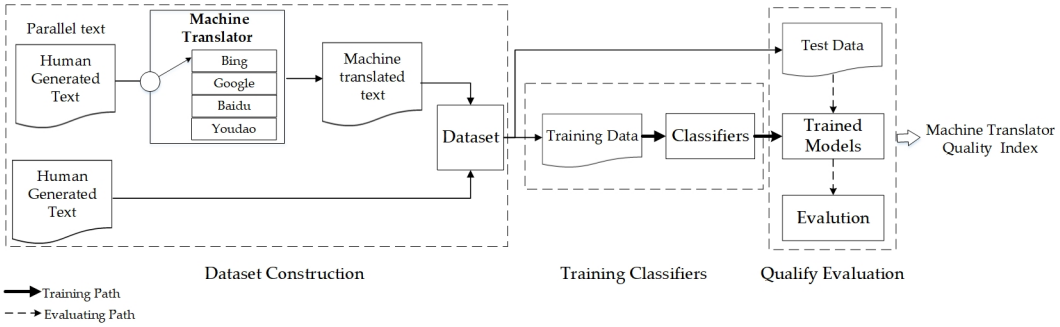

Fig. 1. Framework of the translator evaluation.

3.2 Training of Classifiers

Classifiers are used to assign different observations to different classes based on their features or to predict new data based on currently available data. This is usually done through a training process of applying specific classification rules to historical data. After being trained, the classifier can be used to classify subsequent observations. The classification assignment was to determine how best to apply this label to subsequent data. For each of the objects, the categorization model was trained during the training phase, where the sentences of training data were represented by feature vectors and labelled as '1' or '0', respectively, when training the classifier. '1' indicates that the text is as human translation, and '0' indicates that the text is detected as a machine translator translation.

When we count the word frequency of the words in the dataset, we find that most of the words at the far end of the spectrum are noisy and appear infrequently, which will affect the classification. To avoid the influence of these data on the classification, we only select unigrams and bigrams with frequencies in the top 30000 words as bag-of-words (BOWs) in the training dataset. As shown in Figure 2 and Figure 3, we express each sentence as a feature vector in a one-hot or term-frequency vector representation after extracting the unigrams and bigrams. The feature vector of a text has a positive sign at the indices of unigrams or bigrams found in that text and is zero otherwise. The positive value at the indices of unigrams or bigrams is dependent on whether the type of feature we define is presence or frequency.

- In the presence feature type, mark the position of the word contained in each sentence of English text in the training dataset as '1' in the BOW and mark the other positions as '0' to generate a one-hot vector.

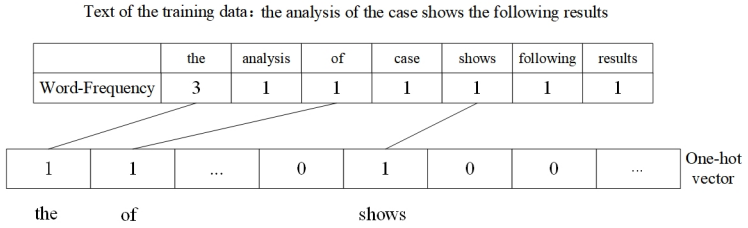

Fig. 2. The one-hot vector representation.

- In the frequency feature type, set the position where the word is contained in each sentence of English text in the training dataset appearing in the BOW as the frequency with which the word appears in the sentence and set it to zero everywhere to generate a term-frequency vector.

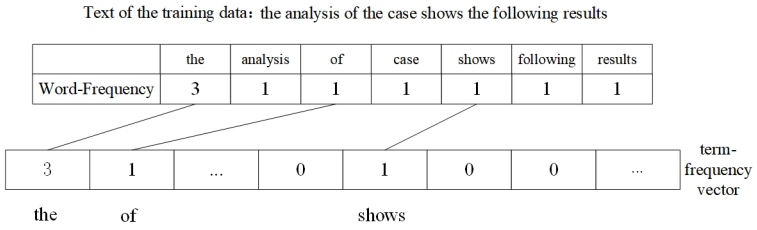

Fig. 3. The term-frequency vector representation.

3.3 Quality Evaluation

Inspired by the utilization of binary classifiers in machine translation detection, the evaluation of machine translators makes it possible to think of it as a binary classification issue, which is the classification of a set of samples into two distinct categories. If the texts generated by machine translators and human-generated texts are indistinguishable, it indicates that the quality of the machine translators is perfect. Rather than employing one binary classifier that outputs one of two labels, we employed five binary classifiers, each labelling the object with a '0' or '1' for their respective object.

Although the above binary classifiers have strong modelling ability and can output stable accuracy, several pretrained classifiers on test datasets predicted different classification results during the experiments, affecting the accuracy of the machine translation quality evaluation. According to a recent trend (An *et al.*, 2020; Chen *et al.*, 2019; Xiao *et al.*, 2018), the majority voting ensemble is most widely employed in a variety of sectors to improve the robustness of the classification model. Majority voting ensembles are a combination strategy for classification problems. This ensemble learning model follows the principle of the minority obeying the majority. It reduces variance through the integration of multiple models; thus, the prediction effect of the majority voting ensemble is better than that of a single classifier. As a result, an ensemble method was used in this study to improve the overall performance of several text categorization models utilizing a majority voting mechanism. The number of classifiers ought to be odd to avoid a tie between forecasted class labels while utilizing the voting procedure. Therefore, in this study, binary classification-based evaluation was performed with five classifiers. The flow chart of the majority voting ensemble is shown in Figure 4.

Through the majority vote ensemble algorithm, the prediction accuracies of the proposed method for four translators are obtained. However, when using translators to generate datasets of translator-translated sentences, the amount of data generated is different due to the difference in the performance of the machine translator, which will affect the results of data analysis. When the dataset is unbalanced, we strive to minimize this type of error. To counteract the effects of these factors, this paper combines the ratio of the training and test sets generated by the same machine translator with classification accuracy to calculate the final quality evaluation index and make their comparability more obvious. We propose a

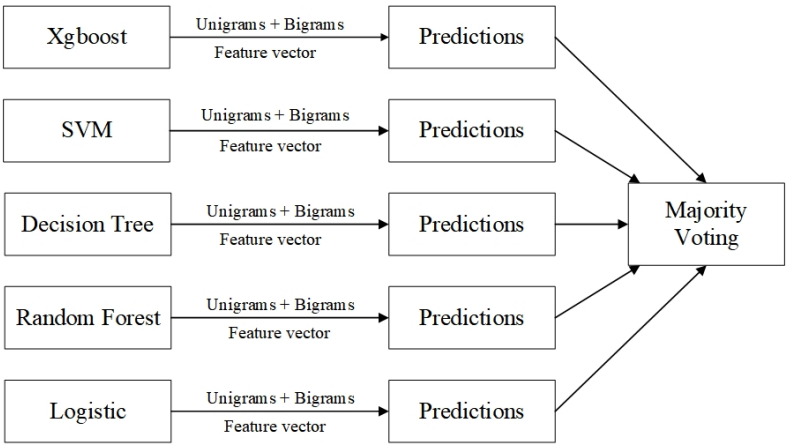

Fig. 4. Flowchart of the majority voting ensemble.

method to calculate the evaluation index by Equation (1):

$$Q = (t/r)^{\frac{TP+TN}{TP+TN+FP+FN}} \tag{1}$$

where r and t are the numbers of training and test datasets, respectively. TN and TP are the true negatives and true positives, respectively, and FN and FP are the false negatives and false-positives, respectively. The evaluation index is mapped to the interval [0,1].

4. Experimental Results

4.1 Dataset Construction

For this study, we collected 9,633 Chinese-English bilingual abstract datasets from the bilingual corpus, which were separated into two distinct datasets: the training and test datasets. The four translators, e.g., the Bing, Google, Baidu, and Youdao translators, were used to translate the Chinese abstracts into English. Due to the unbalanced number of samples belonging to four translator translations, the data were divided in such a way that the test dataset selects the first 1000 paragraphs of the dataset intended to evaluate the translator, and the training dataset selects the remaining 8633 paragraphs of the dataset intended to train classifiers to automatically classify translator-translation and human-translation classes. Then, the paragraphs are divided into sentences. The four translator-related datasets were created for four different translation evaluation projects. Tables 1 and 2 summarize the training and test datasets, respectively. To build a dataset that can be easily studied by various classifiers, raw data must be standardized. To normalize the dataset and reduce its size, we used a large number of preprocessing processes. On the dataset, we performed the following general preprocessing: (1) Strip any punctuation from the words; (2) Convert the words to lower case; (3) Divide paragraphs into sentences.

4.2 Evaluation of the Classifiers

The goal of this experiment was to evaluate machine translators automatically. We perform tests with a variety of different classifiers. To avoid overfitting, we use a test dataset to validate our models. When we combine unigrams and bigrams, we obtain slightly better results than when we solely utilize unigrams. Therefore, only presence with Unigrams + Bigrams (PUB) and frequency with Unigrams + Bigrams (FUB) are used in this experimental configuration, and Table 3 shows the results of the accuracy. It is clear that, in the case of binary classification, different experimental configurations produce different results. To evaluate the quality of the Bing translator, the best accuracy ratios of {XGBoost with PUB, SVM with PUB, DT with FUB, LR with FUB and RF with PUB} are equal to {73.59%, 73.06%, 69.38%, 75.04%, 69.92%}, respectively. To evaluate the quality of the Google translator, the best accuracy ratios

Table 1. Statistics of the training dataset

		Sentence	Unigrams	Bigrams
Total	Google	123755	3505578	3381848
	Bing	107966	3418741	3310775
	Baidu	111140	3320147	3209007
	Youdao	119903	3547164	3427261
Unique	Google	—	36021	573892
	Bing	—	38867	575625
	Baidu	—	32727	536578
	Youdao	—	35775	580849

Table 2. Statistics of the test dataset

		Sentence	Unigrams	Bigrams
Total	Google	13633	398197	384566
	Bing	11430	380015	368585
	Baidu	12625	394509	381884
	Youdao	12852	396263	383411
Unique	Google	—	12939	119353
	Bing	—	13145	116659
	Baidu	—	12064	115224
	Youdao	—	12891	120434

of {XGBoost with PUB, SVM with PUB, DT with FUB, LR with PUB and RF with PUB} are {69.21%, 69.33%, 64.39%, 69.15%, 65.13%}, respectively. To evaluate the quality of the Baidu translator, the best accuracy ratios of {XGBoost with PUB, SVM with PUB, DT with FUB, LR with PUB and RF with FUB} are {66.50%, 68.24%, 64.35%, 68.94%, 63.13%}, respectively. To evaluate the quality of the Youdao translator, the best accuracy ratios of {XGBoost with PUB, SVM with PUB, DT with FUB, LR with FUB and RF with PUB} are {60.75%, 60.69%, 59.55%, 60.76%, 55.95%}, respectively. There is a certain gap in classification accuracy compared with previous methods of sentence-by-sentence comparison of reference translations. Our method does not need reference translations and relies only on the machine learning experience of binary classifiers for classification. However, in the quality evaluation of the machine translators, our method only requires the monolingual machine translation output as input and is obviously superior to other evaluation methods in algorithm complexity.

Although the classification models selected in this paper have strong modelling ability and can output stable accuracy, different classification models have different classification decisions due to their different internal structures; that is, the same text may have different judgement results in different classifiers. To further improve accuracy, the majority voting ensemble of forecasts from the five binary classification models is then used. Table 4 shows the accuracies of each of these machine translators as well as their majority voting ensemble.

4.3 Evaluation of the Classifiers

We report the quality evaluation index of each machine translator in Table 5. We count the numbers of TP, TN, FP, and FN after majority voting. TP and TN reflect the effectiveness of this method in evaluating machine translators. FP and FN reflect the difference between the translation quality of the machine translator and manual translation. In our experiment, all classifiers are shown to be effective and achieve the best performance. Finally, Equation (1) is used to compute the quality evaluation index of the Bing, Google, Baidu and Youdao translators, which are 0.1695, 0.1939, 0.2076 and 0.2557, respectively,

Table 3. Statistics of the test dataset

Configuration	Machine Translator			
	Bing	Google	Baidu	Youdao
Presence with Unigrams + Bigrams				
XGBoost	**73.59**	**69.21**	**66.50**	**60.75**
SVM	**73.06**	**69.33**	**68.24**	**60.69**
DT	68.26	62.21	63.58	59.19
LR	73.79	**69.15**	**68.94**	60.29
RF	**69.92**	**65.13**	63.07	**55.95**
Frequency with Unigrams + Bigrams				
XGBoost	73.46	69.07	66.41	60.50
SVM	72.81	68.39	67.32	58.97
DT	**69.38**	**64.39**	**64.35**	**59.55**
LR	**75.04**	68.99	68.90	**60.76**
RF	69.72	64.80	**63.13**	55.63

Table 4. Statistics of the test dataset

Machine Translator	Majority Voting Ensemble
Google	71.01
Bing	**75.65**
Baidu	69.37
Youdao	60.84

objectively reflecting that the translation quality of the Youdao translator is obviously better than those of the other three machine translators. These results indicate that our method can be used to evaluate the quality of machine translators.

Table 5. Statistics of the test dataset

Machine Translator	Quality Evaluation Index
Google	0.1939
Bing	0.1695
Baidu	0.2076
Youdao	**0.2557**

5. Discussion

In this work, we presented a strategy of machine translator evaluation without references that includes training classifiers and a better evaluation index design. The resulting classifiers provide a new state of the art for evaluating the quality of machine translators. It was shown that the approach without references can also be applied to evaluate machine translators. Moreover, in comparison with other commonly used methods, it was shown that it exhibits preferable performance.

References

Alexandra, A., & **Alexey, M. (2011)**. Building a web-based parallel corpus and filtering out machine translated text. In *Proceedings of the Workshop on Building and Using Comparable Corpora* (pp. 136–144).

Ayala, B.R. & **Chen, J.P. (2017)**. Evaluating Translation Quality via Utilizing Skopos Theory. *Proceedings of the 2016 International Conference on Education, Management, Computer and Society*, (pp. 231–232).

Anurag, S., Malik, P., & **Mrudula, Y. (2020)**. Statistical Analysis of Machine Translation Evaluation Systems for English- Hindi Language Pair. *Recent Advances in Computer Science and Communications*, **13**(5), 864–870.

An, N., Ding, H., Yang, J., Au, R., & **Ang, T.F. (2020)**. Deep ensemble learning for Alzheimer's disease classification. *J. Biomed. Inf*, **105**, 103411.

Banerjee, S., & **Lavie, A. (2005)**. 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 Machine Translation and/or Summarization* (pp. 65–72).

Barzilay, B., & **Lapata, M. (2005)**. Modeling local coherence: An entity-based approach. Computational Linguistics, 141–148.

Chung, H. Y. (2020). Automatische Evaluation der Humanübersetzung: BLEU vs. METEOR. *Lebende Sprachen*, **65**(1), 181–205.

Comelles, E., Gimenez, J., Marquez, L., & **Castellon, I. (2010)**. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics- MATR* (pp. 333–338).

Cai, N. & **Zhou, J. (2016)**. A machine learning approach to evaluating translation quality. *Digital Libraries*, (pp. 281–282).

Cindy, A.M., & **Martin, J.P. (2020)**. Dataset for comparable evaluation of machine translation between 11 South African languages. Data in Brief, 29(C), 105–146.

Coppersmith, D., Hong, S.J., & **Hosking, J.R.M. (1999)**. Partitioning nominal attributes in decision tree. *Data Mining and Knowledge Discovery*, **3**, 197–217.

Chen, T., & **Guestrin, C. (2016)**. XGBoost: a Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). USA : San Francisco California.

Chen, Y., Wang, Y., Gu, Y., He, X., Ghamisi, P., & **jia, X. (2019)**. Deep learning ensemble for hyperspectral image classification. *IEEE J. Select. Topic. Appl. Earth Observ. Rem. Sens*, **12**(6) 1882-1897.

Duan, H., Wang, L., & **Zhang, C. (2020)**. Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions. *RSC Adv*, **10**(3), 1371–1378.

Ehud, R. W. (2018). A Structured Review of the Validity of BLEU. Computational Linguistics, 44(3), 393–401.

Grosz, B. J., Joshi, A. K., & **Weinstein, S. (1995)**. Centering: A framework for modeling the local coherenceof discourse. *Computational Linguistics*, **21**(2), 203–226.

GUO, X.J., ZHANG, C.C., & **LUO, W.R. (2020)**. Urban impervious surface extraction based on multi-features and random forest. *IEEE Access*, **8**, 226609–226623.

Hardmei, C., & **Federico, M. (2010)**. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation* (pp. 283–289).

Iida, R., & **Tokunaga, T. (2012)**. A metric for evaluating discourse coherence based on coreference resolution. In *Proceedings of COLING 2012: Posters* (pp. 483–494).

Joty, S.,Guzmán, F., & **Màrquez, L.(2014)**. Disco TK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 402–408).

Keiji, Y., & **Sugaya F. (2001)**. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. In *Proceedings of the MT Summit Conference* (pp. 373–378).Santiago de Compostela.

Kexin, Y. (2020). An automatic evaluation metric for Ancient-Modern Chinese translation., 33(8), 1–13.

Karamanis, N., Poesio, M., Mellish, C., & **Oberlander, J. (2004)**. Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Evaluating centering-based metrics of coherence using a reliably annotated corpus.* (pp. 391–398).

Khan, W. and Daud, A. and Nasir, J.A., & **Amjad, T. (2016)**. A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait Journal of Science*, **43**(4), 95–113.

Khan, K. and Ullah, A., & **Baharudin, B. (2016)**. Pattern and semantic analysis to improve unsupervised techniques for opinion target identification. *Kuwait Journal of Science*, **43**(1), 129–149.

Lin, Z., & **Kan, M. Y. (2011)**. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics* (pp. 997–1006).

Matthew, S., Bonnie, D., Richard, S., Linnea, M., & **John, M. (2006)**. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (pp. 223–231). America.

Matthew, G. S. (2010). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, **23**(2-3), 117–127.

Mann, W. C., & **Thompson, S. A. (1988)**. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3), 243–281.

Munkova, D. (2018). Detecting errors in machine translation using residuals and metrics of automatic evaluation. *Robotics Machine Learning*, (pp. 148–149).

Melamed, I.D., Freen, R., & **Turian, J.P. (2003)**. TurianJ.P. Precision and recall of machine translation. In *Proceedings of the NAACL/Human Language Technology*. Canada: Edmonton.

Ogunleye, A., & **Wang, Q.G. (2018)**. Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease. In *Proceedings of the 2018 IEEE 14th International Conference on Control and Automation (ICCA)* (pp. 805–810). Anchorage, AK, USA, **12–15** .

Papineni, K., Roukos, S.,Ward, T., & **Zhu, W. J. (2002)**. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Rawat, P., Kumar, S., & **Michael, G.M. (2018)**. An in-silico method for identifying aggregation rate enhancer and mitigator mutations in proteins. *Int J Biol Macromol*, **118**(Pt A), 1157–1167.

Reed, P., & **Wu, Y. (2013)**. Logistic regression for risk factor modelling in stuttering research. *Journal of Fluency Disorders*, **38**, 88–101.

Spencer, R.,Will, L., & **Chris, Q. (2011)**. MT detection in web-scraped parallel corpora. In *Proceedings of the Machine Translation Summit (MT Summit XIII)*.

Shimanaka, H., Kajiwara, T., & **Komachi, M. (2019)**. IEICE Transactions on Fundamentals of Electronics. Communications and Computer Sciences.

Sen, S., & **Raghunathan, A. (2018)**. Approximate Computing for Long Short Term Memory (LSTM) Neural Networks. *IEEE Transactions on Computer Aided Design of Integrated Circuits* & *Systems*, **37**(11), 2266–2276.

Wojciech, D., Jakub, N., & **Michal, K. (2021)**. Evolving data-adaptive support vector machines for binary classification. Knowledge-Based Systems, 227.

Wang, J., Li, L., Yang, P., Chen, Y., Zhu, Y., Tong, M., & **Li, X. (2018)**. Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine. *Lasers Med Sci*, **33**(6), 1381–1386.

Wong, B., & **Kit, C. (2011)**. Extending machine translation evaluation metrics with lexical cohesion todocument level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics* (pp. 1060–1068).

Xiao, Y., Wu, J., Lin, Z., & **Zhao, X. (2018)**. A deep learning-based multi-model ensemble method for cancer prediction, Comput. *Methods Progr. Biomed*, **153**, 1–9.

YAO, J.M. and Qu, Y.Q. and Zhu, Q.M. & **Zhang, J. (2006)**. A Visualization method for machine translation evaluation results. In *Proceedings of the 20th Asia-pacific International Conference on Language, Information and Computing* (pp. 401–404).

Yasuhiro, A., Imamura, K., & **Sumita, E. (2001)**. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of the MT Summit Conference* (pp. 15–20).Santiago de Compostela.

Young, K.S., & **Arun, U. (2021)**. Majority voting ensemble with a decision trees for business failure prediction during economic downturns. Journal of Innovation & Knowledge.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & **Si, Y. (2018)**. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, **6**, 21020–21031.

Zhang, W., Yoshida, T., & **Tang, X. (2008)**. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, **21**, 879–886.

Submitted :	18/03/2022
Revised:	19/05/2022
Accepted:	23/05/2022
DOI:	10.48129/kjs.splml.19547