NoDeeLe: A Novel Deep Learning Schema for Evaluating Neural Machine Translation Systems

Despoina Mouratidis[1][0000−0002−2844−5488], Maria Stasimioti[2][0000−0001−9541−4676], Vilelmini Sosoni[2][0000−0002−9583−4651], and Katia Lida Kermanidis[1][0000−0002−3270−5078]

[1] Department of Informatics, Ionian University, 491 00 Corfu, Greece {c12mour,kerman}@ionio.gr

[2] Department of Foreign Languages, Translation and Interpreting, Ionian University, 491 00 Corfu, Greece {stasimioti,sosoni}@ionio.gr

Abstract. Due to the wide-spread development of Machine Translation (MT) systems—especially Neural Machine Translation (NMT) systems—MT evaluation, both automatic and human, has become more and more important as it helps us establish how MT systems perform. Yet, automatic evaluation metrics have lagged behind, as the most popular choices (e.g., BLEU, METEOR and ROUGE) may correlate poorly with human judgments. This paper seeks to put to the test an evaluation model based on a novel deep learning schema (NoDeeLe) used to compare two NMT systems on four different text genres, i.e. medical, legal, marketing and literary in the English-Greek language pair. The model utilizes information from the source segments, the MT outputs and the reference translation, as well as the automatic metrics BLEU, METEOR and WER. The proposed schema achieves a strong correlation with human judgment (78% average accuracy for the four texts with the highest accuracy, i.e. 85%, observed in the case of the marketing text), while it outperforms classic machine learning algorithms and automatic metrics.

Keywords: Machine Learning · Deep Learning Schema · Neural Machine Translation · Pairwise Evaluation.

1 Introduction

Recently, studies in Natural Language Proccessing (NLP) have been using neural networks [31,1]. Neural networks have made significant progress in several NLP tasks including MT [20], summarization [7], dialogue generation [21] and image captioning [11]. The evaluation of MT systems is a crucial field of research, as has been highlighted by a number of researchers [34,15,16,3], given that it is used to compare different systems but also to identify a system's weaknesses and help improve it. Various methods have been suggested for the evaluation of MT—both automatic and human [4]. Although, human evaluation is considered to be the best indicator of a system's quality, it is an expensive and time-consuming process, so it cannot be readily used for the development of the

Translation and Interpreting Technology Online, pages 37–47,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_005

MT system. As a result, MT researchers and developers mostly use automatic evaluation metrics which constitute an acceptable estimation quality and they are easy and cheap to compute. Some of them rely on score-based metrics, such as Bilingual Evaluation Understudy (BLEU) [26], National Institute of Standards and Technology (NIST) [12] and Word Error Rate (WER) [30], metrics using external resources, like METEOR [10], and neural metrics such as ReVal [17] and Regressor Using Sentence Embeddings (RUSE) [28], while some others use machine learning schemata [13,32,23]. Automatic evaluation methods must be evaluated with specific criteria. According to Banerjee and Lavie [2], a satisfactory automated evaluation system should meet the following conditions: high correlation with human judgments quantified in relation to translation quality, sensitivity to nuances in quality among systems or outputs of the same system in different stages of its development, result consistency, reliability, a great range of fields and speed and usability. The most important condition is considered to be correlation with human judgment [29]. Yet, the automatic evaluation metrics mentioned above have lagged behind, as they do not correlate well with human judgments [27].

This paper seeks to put to the test an evaluation model based on a novel deep learning schema developed by Mouratidis et al. [25] used to compare two NMT systems on four different text genres, i.e. medical, legal, marketing and literary in the English-Greek language pair. The model, NoDeeLe, utilizes information from the source segments, the MT outputs and the reference translation, as well as the automatic metrics BLEU, METEOR and WER.

2 Related Work

Deep Learning (DL) is one of the fastest-growing fields of Information Technology (IT) today being used among others for MT evaluation. Duh [13] decomposes rankings into parallel decisions, with the best translation for each candidate pair predicted, using a ranking-specific feature set, BLEU score and the Support Vector Machine (SVM) classifier. A similar pairwise approach was proposed by Mouratidis and Kermanidis [22], using a random forest (RF) classifier. Cho et al. [6] proposed a score-based schema to learn the translation probability of a source phrase to a target phrase (MT output) with a Recurrent Neural Network (RNN) encoder-decoder. They showed that this learning schema has improved the translation performance. The schema proposed by Sutskever et al. [31] is similar to the work by Cho et al. [6], but Sutskever et al. chose the top 1000 best candidate translations produced by a Statistical Machine Translation (SMT) system with a Long Short-Term Memory (LSTM) sequence-to-sequence model. Wu et al. [32] also trained a deep LSTM network to optimize BLEU scores focusing on German-English and German-French language pairs, but they found that the improvement in BLEU scores did not reflect the human evaluation of translation quality. Mouratidis et al. [24] used LSTM layers in a learning framework for evaluating pairwise MT outputs using vector representations, in order to show that the linguistic features of the source text (ST) can affect MT evaluation.

Gehring et al. [14] proposed an architecture for sequence to sequence modeling based on a Convolutional Neural Network (CNN). The model is equipped with linear units [9] and residual connections [18].

3 Materials and Methods

3.1 Dataset

The STs used in this study are four texts of comparable complexity, i.e. with a Lexile score between 1210 and 1400, belonging to different genres: medical (*T1*), legal (*T2*), literary (*T3*) and marketing (*T4*). All texts were originally written in English. The medical text is a 382-word excerpt from a clinical trial retrieved from the National Center for Biotechnology Information, the legal text is a 367-word excerpt from a purchase agreement, the literary text is a 365-word excerpt from the book *The English* by Jeremy Paxman, while the marketing text is a 410-word excerpt about the Venice Simplon-Orient-Express holidays retrieved from the website of luxury travel tour operator The Luxury Holiday Company. The Lexile score was calculated on the basis of the Lexile Analyzer[1] which relies on an algorithm to evaluate the reading demand –or text complexity– of books, articles, and other materials. In particular, it measures the complexity of the text by breaking down the entire piece and studying its characteristics, such as sentence length and word frequency, which represent the syntactic and semantic challenges that the text presents to a reader.

The STs were machine-translated without any pre-editing and the NMT systems used to produce the raw MT output were DeepL[2] and Google Translate[3] (output obtained June 2, 2021). Google Translate and DeepL are both generic NMT systems that use state-of-the-art AI to translate texts from one language into another. However, these systems differ in the technology they use and the language data they are trained on. More specifically, DeepL uses CNNs and is trained on the Linguee bilingual corpora database, while Google Translate, uses RNNs and is trained on various digital resources in many languages [35].

The reference translations, i.e. the gold-standard human translations, were produced by highly experienced professional translators. In particular, the medical text was translated by a professional translator specialising in the Life Sciences with over 15 years of experience, the legal text was translated by a professional translator and Law graduate with over 10 years of translation experience, while the literary and marketing texts were translated by a professional translator specialising in creative genres and having more than 20 years of experience.

3.2 The Feature Set Used

Two different features categories were employed from source segments, MT outputs and reference translation.

[1] https://lexile.com/
[2] https://www.deepl.com/translator
[3] https://translate.google.com/

The first one derives from string-based linguistic features and the second one from MT evaluation automatic metrics. The first category contains i. string-based similarity features (such as length in words and characters, the longest word length, some ratios e.g. the ratio between lengths in words in the source segments and the two MT outputs, the ratio between longest words from source segments and the two MT outputs and reference translation, etc., the percentage of segments similarity, suffix similarity etc.) and ii. noise features (such as repeated words or special characters). All the features were calculated for the two MT outputs, the source segments and the reference translation. More details on the feature set used can be found in Mouratidis et al. [24]. The second category contains the BLEU score, METEOR and WER.

3.3 Word Embeddings

Word embeddings helped us to model the relations between the two MT outputs and the reference translation. In this paper, the embedding layer, the one provided by Keras [19], is used for the two MT outputs and the reference translation. The encoding function applied is the one-hot function. The embedding layer size, in number of nodes, is 16.

3.4 The DL Architecture

The deep learning schema in Figure 1 is used for classification purposes.

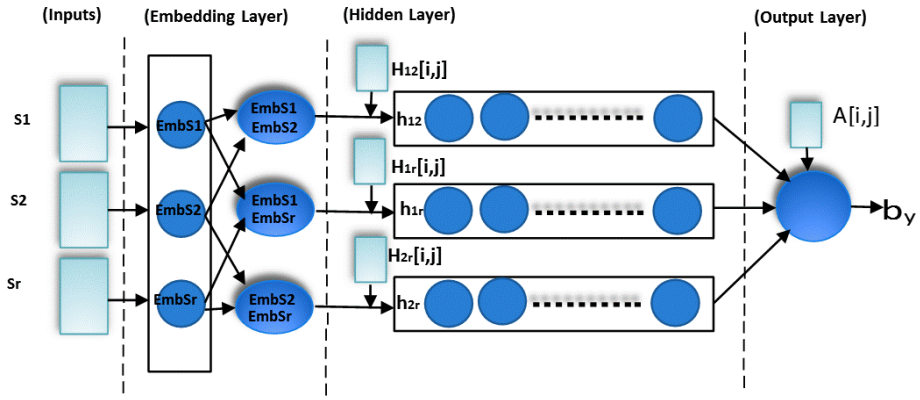

Fig. 1. Deep learning architecture

The input segments in the learning schema are the two MT outputs *S1*, *S2* and the reference translation *Sr*. These segments are converted into numerical vectors (*EmbS1*, *EmbS2*, *EmbSr*) by passing the embedding layer and then they are merged by pair in order to become the input to the hidden layers. In this step, the architecture takes as an extra input the matrices H containing linguistic

features from the source segments and the automatic metrics BLEU, METEOR and WER. The architecture takes an extra input to the output layer, the matrix A containing the linguistic features from the MT outputs and the reference translation. Finally, we used as ground truth the ranking information produced by two linguists —both native speakers of Greek, both translators with over 10 years of professional experience each and with a specialisation in MT evaluation/annotation in the English-Greek language pair. The linguists ranked the MT outputs of the four texts at sentence level as follows: 1 if the DeepL MT output is better than the Google MT output, and 0 if the Google MT output is better than the DeepL MT output. The inter-annotator agreement was calculated using Cohen's kappa coefficient (κ) which measures the inter-annotators' reliability; this can take a value between 0 and 1 where 1 indicates perfect agreement and 0 indicates no agreement [8]. In the cases of disagreement between the two annotators, a third, mediating annotator was introduced to resolve the disagreement [33]. The mediating annotator was a professional translator with 15 years of translation experience in the English-Greek language pair and extensive experience in MT evaluation/annotation. The network model architecture for the experiments is a classic architecture of LSTM and feedforward layers.

To avoid over-fitting, a dropout rate of 0.05 is applied, using the binary cross entropy as a loss function and 10-fold Cross Validation. More details about the model's parameters can be found in [25].

4 Results

According to the annotators, and as it emerges from Figure 2, DeepL performed better than Google Translate for all texts. It should be noted that an almost perfect agreement between the annotators for all four texts (κ=0.83 for *T1*, κ=1.0 for *T2*, κ=0.92 for *T3* and κ=0.85 for *T4*) was observed. In the few cases of disagreement, the mediating annotator's decision was used.

Unequal values between the classes were observed with the class belonging to Google Translate being the minority class. The SMOTE supervised filter [5] was applied to the minority class. Figure 3 presents the classification results (classification accuracy) for the two MT outputs over the four different datasets. It emerges that the classification accuracy level is higher in the case of the marketing text followed by the legal and the literary text. The lowest accuracy level is observed in the case of the medical text, most probably due to its rich and highly-specialised terminology. We also applied a SMOTE filter with a view to improving the model accuracy. Indeed, an increase of 2% of the classification accuracy for the medical text, 5% for the legal and the marketing text, and 3% for the literary text is observed. The above accuracy results are in accordance with the annotators' results (see Figure 2). Better accuracy results (*F1 score*) are observed for DeepL (*S1*) compared to Google Translate (*S2*) for all texts (Figure 4).

The BLEU and METEOR scores for the MT outputs of the four texts are given in Figure 5. In particular, the medical text received the highest score

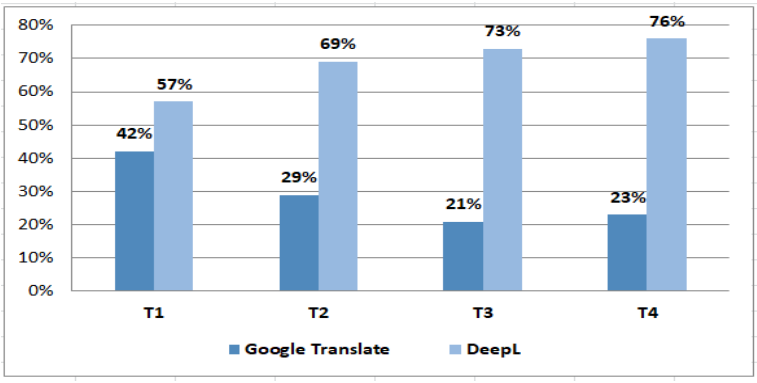

Fig. 2. Ranking information

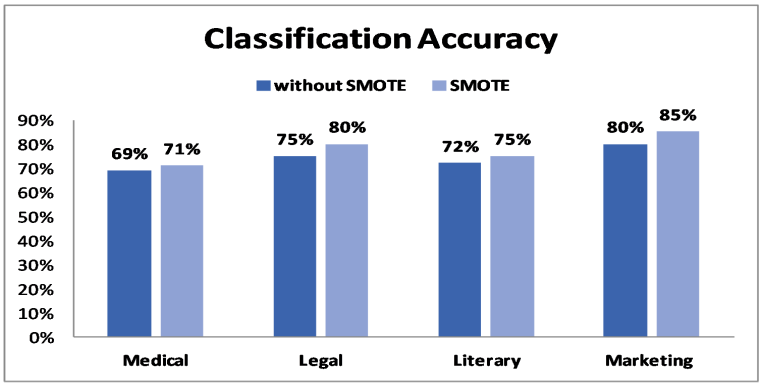

Fig. 3. Accuracy performance with and without SMOTE filter

for both metrics followed by the legal text, while the literary text received the lowest score. Interestingly, the scores are not in line with the results of NoDeeLe, i.e. the proposed deep learning schema, according to which the medical text received the worst classification accuracy and the marketing text received the best classification accuracy. In addition, although DeepL (*S1*) performed better in all cases according to NoDeeLe, Google Translate (*S2*) performed better in the case of the literary and marketing text according to BLEU, and in the case of the medical text according to METEOR. As far as the legal text is concerned, no difference was observed between the automatic metrics and NoDeeLe.

Apart from BLEU and METEOR, NoDeeLe was also compared to other methods. For that reason, additional experiments were carried out using different classifiers e.g. SVM and RF using the WEKA framework as backend. The SVM and the RF classifiers were trained on the same data and feature set as NoDeLee. As depicted in Figure 6, NoDeeLe achieves stronger correlation with the human judgments (78% average accuracy for the four texts), compared to

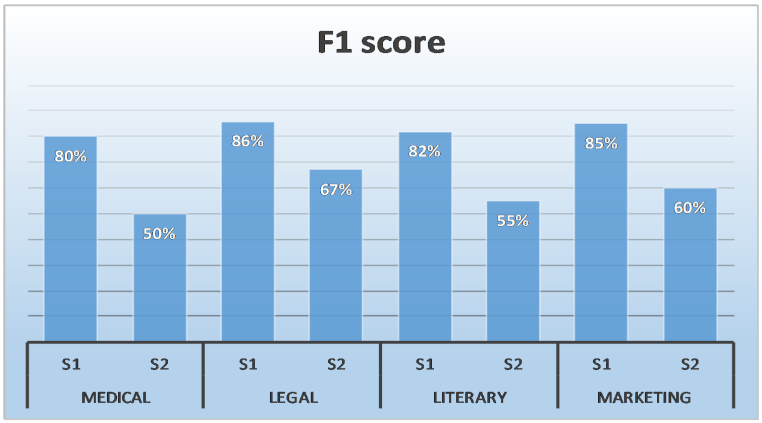

Fig. 4. F1 score per system and per text genre

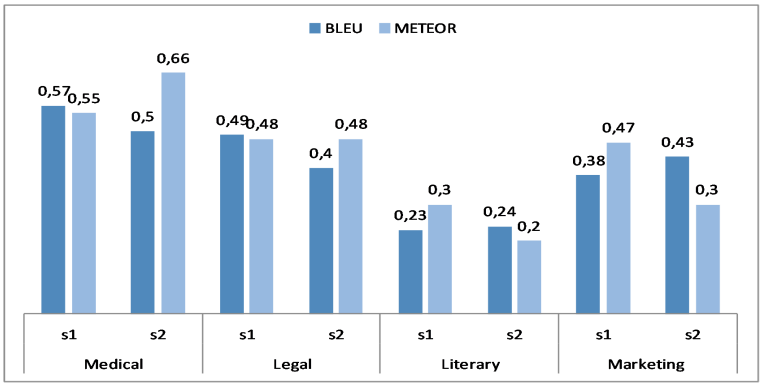

Fig. 5. Automatic Metrics BLEU and METEOR

the RF classifier (74% average accuracy for the four texts) and the SVM classifier (67% average accuracy for the four texts). Unlike BLEU and METEOR scores, NoDeeLe as well as the RF and SVM classifiers indicate that the marketing text has the best classification accuracy, followed by the legal text and the literary text, while the medical text has the worst classification accuracy. In addition, NoDeeLe as well as the RF and SVM classifiers reveal that DeepL (S1) performed better in all text genres in contrast with BLEU and METEOR, with the former showing that Google Translate (*S2*) performed better in the case of the literary and marketing text, and the latter showing that it performed better in the case of the medical text.

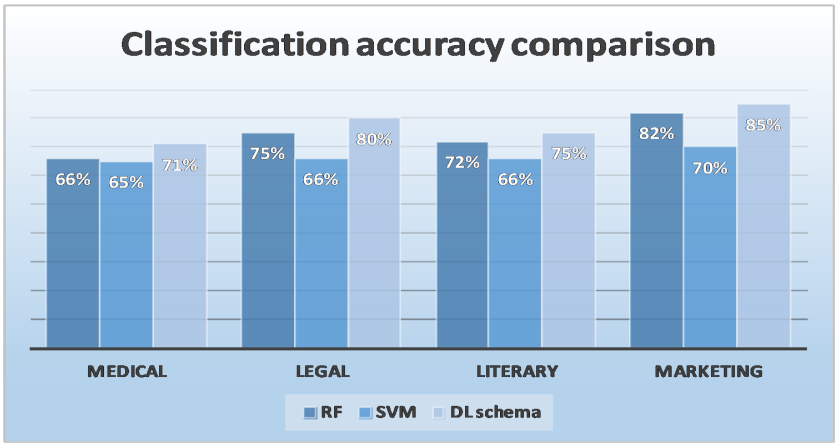

Fig. 6. Classification accuracy comparison with other algorithms

5 Conclusions and Future Work

In this paper, a deep learning novel schema for evaluating NMT systems and outputs is presented and discussed. The schema, i.e. NoDeeLe, used information from the source segments, the MT outputs and the reference translations as well as automatic metrics, and was applied in four different text genres: medical, legal, literary and marketing. Experimental results showed that NoDeeLe achieves stronger correlation with the human judgments compared to the RF classifier and the SVM classifier. Unlike BLEU and METEOR, NoDeeLe, as well as the RF and SVM classifiers, indicate *i.* that the marketing text has the best classification accuracy and the medical text the worst classification accuracy and *ii.* DeepL (*S1*) performed better in all text genres. These findings suggest that the BLEU and METEOR automatic metrics may not be appropriate for the evaluation of NMT output, as has been also indicated by other studies [4].

To complement and expand this study, we aim to explore if pre-trained embeddings e.g. fasttext, could improve classification accuracy, especially concerning texts with specialised terminology. In addition, we are planning to test: i. another neural network structure and ii. a learned evaluation metric, the BLEURT metric [27], on the same datasets. Finally, in order to further explore the observed difference between the BLEU and METEOR automatic metrics and NoDeeLe, we are planning to carry out a refined human error analysis to evaluate the linguistic quality of the MT outputs.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

2. 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 Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://www.aclweb.org/anthology/W05-0909
3. Bentivogli, L., Cettolo, M., Federico, M., Christian, F.: Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment (2018)
4. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering **26**(2), 137–161 (2020)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP (2014)
7. Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–98 (2016)
8. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
9. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International conference on machine learning. pp. 933–941. PMLR (2017)
10. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
11. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL-IJCNLP 2015. pp. 100–105. Association for Computational Linguistics (ACL) (2015)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. pp. 138–145 (2002)
13. Duh, K.: Ranking vs. regression in machine translation evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 191–194 (2008)
14. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning. pp. 1243–1252. PMLR (2017)
15. Giménez, J., i Villodre, L.M.: Asiya: An open toolkit for automatic machine translation (meta-)evaluation. In: Prague Bull. Math. Linguistics (2010)
16. GRAHAM, Y., BALDWIN, T., MOFFAT, A., ZOBEL, J.: Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering **23**(1), 3–30 (2017). https://doi.org/10.1017/S1351324915000339
17. Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P.: Deep learning with limited numerical precision. In: International Conference on Machine Learning. pp. 1737–1746 (2015)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Keras, K.: Deep learning library for theano and tensorflow. 2015 (2019), https://keras.io/
20. Koehn, P.: Statistical machine translation. Cambridge University Press (2009)
21. Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J.: Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1192–1202 (2016)
22. Mouratidis, D., Kermanidis, K.L.: Automatic selection of parallel data for machine translation. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 146–156. Springer (2018)
23. Mouratidis, D., Kermanidis, K.L.: Ensemble and deep learning for language-independent automatic selection of parallel data. Algorithms **12**(1), 26 (2019)
24. Mouratidis, D., Kermanidis, K.L., Sosoni, V.: Innovative deep neural network fusion for pairwise translation evaluation. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 76–87. Springer (2020)
25. Mouratidis, D., Kermanidis, K.L., Sosoni, V.: Innovatively fused deep learning with limited noisy data for evaluating translations from poor into rich morphology. Applied Sciences **11**(2), 639 (2021)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
27. Sellam, T., Das, D., Parikh, A.: Bleurt: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892 (2020)
28. Shimanaka, H., Kajiwara, T., Komachi, M.: Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 751–758 (2018)
29. Specia, L., Raj, D., Turchi, M.: Machine translation evaluation versus quality estimation. Machine Translation **24**(1), 39–50 (2010). https://doi.org/10.1007/s10590-010-9077-2, https://doi.org/10.1007/s10590-010-9077-2
30. Su, K.Y., Wu, M.W., Chang, J.S.: A new quantitative quality measure for machine translation systems. In: COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics (1992)
31. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
32. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
33. Zhang, Z., Chapman, S., Ciravegna, F.: A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality. In: Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses. p. 301–315. EKAW'10, Springer-Verlag, Berlin, Heidelberg (2010)
34. Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., Zhao, T.: Diagnostic evaluation of machine translation systems using automatically constructed linguistic checkpoints. In: Proceedings of the 22nd International Conference on Computational

Linguistics (Coling 2008). pp. 1121–1128. Coling 2008 Organizing Committee, Manchester, UK (Aug 2008), https://www.aclweb.org/anthology/C08-1141
35. Ziganshina, L.E., Yudina, E.V., Gabdrakhmanov, A.I., Ried, J.: Assessing human post-editing efforts to compare the performance of three machine translation engines for english to russian translation of cochrane plain language health information: Results of a randomised comparison. In: Informatics. vol. 8, p. 9. Multidisciplinary Digital Publishing Institute (2021)