Innovative Deep Neural Network Fusion for Pairwise Translation Evaluation

Despoina Mouratidis, Katia Kermanidis, Vilelmini Sosoni

Innovative Deep Neural Network Fusion for Pairwise Translation Evaluation

Despoina Mouratidis [1][0000-0002-2844-5488], Katia Lida Kermanidis [1][0000-0002-3270-5078] and Vilelmini Sosoni [2][000-0002-9583-4651]

[1] Department of Informatics, Ionian University, Tsirigoti Squ. 7, 49100 Corfu, Greece
[2] Department of Foreign Languages, Translation and Interpreting, Ionian University, Tsirigoti Square 7, 49100 Corfu, Greece
{c12mour,kerman,sosoni}@ionio.gr

Abstract. A language independent deep learning (DL) architecture for machine translation (MT) evaluation is presented. This DL architecture aims at the best choice between two MT (*S1, S2*) outputs, based on the reference translation (*Sr*) and the annotation score. The outputs were generated from a statistical machine translation (SMT) system and a neural machine translation (NMT) system. The model applied in two language pairs: English - Greek (EN-EL) and English - Italian (EN-IT). In this paper, a variety of experiments with different parameter configurations is presented. Moreover, linguistic features, embeddings representation and natural language processing (NLP) metrics (BLEU, METEOR, TER, WER) were tested. The best score was achieved when the proposed model used source segments (*SSE*) information and the NLP metrics set. Classification accuracy has increased up to 5% (compared to previous related work) and reached quite satisfactory results for the Kendall τ score.

Keywords: Machine learning, machine translation evaluation, deep learning, neural network architecture, pairwise classification.

1 Introduction

Deep neural networks are demonstrating a large impact on NLP. NMT [2, 14, 28, 26], in particular, has gained increasing popularity since it has shown remarkable results in several tasks and its effective approach has had a strong influence on other related NLP tasks, such as dialogue generation [8].

The evaluation of MT systems is a vital field of research, both for determining the effectiveness of existing MT systems (evaluation of the classification performance) and for guiding the MT systems modeling. Progress in the field of MT relies on assessing the quality of a new system through systematic evaluation, such that the new system can be shown to perform better than pre-existing systems. The difficulty arises in the definition of a better system. When assessing the quality of a translation, there is no single correct answer; rather, there may be any number of possible correct translations. In addition, when two translations are only partially correct -but in different ways- it is difficult to distinguish quality.

Many methods for MT evaluation have been employed. There are metrics that focus on the MT output evaluation, such as BLEU [18], METEOR [4], TER [24] and WER [25]. BLEU score is maybe the most famous and widely-used metric in MT evaluation. The closer an MT output is to the professional translation, the higher the BLEU score is. The BLEU score suffers from several shortcomings i.e. it doesn't handle morphologically rich languages well and it doesn't map well to human judgements. Several other metrics, that address these issues, are used, such as METEOR. The METEOR score has a good correlation with human judgement at the segment level. It is based on the alignment between the MT outputs and the professional translation. Alignments are based on synonym and paraphrase matches between words and phrases. The translation error rate (TER) and word error rate (WER) are other commonly-used metrics. They are based on the matching of the MT outputs with the professional translation. They measure the minimum number of edits needed to change the original output translation into the professional translation. Other metrics focus on performance evaluation. In some studies ([15], [17]), parallel corpora are used and showed that certain string-based features, e.g. the length of the segments, and similarity-based features e.g. the ratio of common suffixes shared between the MT outputs and the reference, could improve the MT system performance. They considered the task as a classification problem and they used Random Forest (RF) as classifier.
NMT can potentially perform end-to-end translation, though many NMT systems are still relying on language-dependent pre- and post-processors, which have been used in traditional SMT systems. Moses [11], a toolkit for SMT, implements a reasonably useful pre- and post-processor. A language dependent processing also makes it hard to train multilingual NMT models.
It is important for the NLP community to develop a simple, efficient and language independent framework for automatic MT evaluation. A few studies have been reported using learning frameworks. Duh (2008) [5] uses a framework for ranking translations in parallel settings, given information of translation outputs and a reference translation. This study showed that ranking achieves higher correlation to human judgments when the framework makes use of a ranking specific feature set and of BLEU score information. They have tested the framework performance using Support Vector Machine (SVM). Another important work is presented by [7] who used syntactic and semantic information about the reference and the machine-generated translation as well, by using pre-trained embeddings and the BLEU translations scores. They used a feedforward neural network (NN) to decide which of the MT outputs is better. A learning scheme to classify machine-generated translations using information from numerous linguistic features and hand-crafted word embeddings from two MT outputs and one reference translation is presented from [16]. They used a convolutional NN to choose the right translation among two provided.
In this paper, we introduce a learning schema, for evaluating MT, similar to that of a preliminary study [16], but we extend it to a new level, both in terms of number of feature and their representation and learning framework as well.
Compared to that study, the present approach includes the following novelties:

- the utilization of a deeper NN architecture. More hidden layers and different types were tested (Dense and LSTM layers).

- the inclusion of an NLP metric set (BLEU score, METEOR score, TER, WER).
- the use of the linguistic information from the *SSE* in EN. 18 string-based features were calculated and used as an extra input to the DL architecture.
- the accuracy exploration of different inputs to the hidden layers (the NLP set and the string-based features).

To the best of the authors' knowledge, this is the first time that information of the *SSE* combined with handcrafted features, embeddings and a set of NLP metrics are used from a DL architecture for a classification task.

2 Materials and Methods

The current section presents the corpora, the features and NLP set as well as the DL architecture used in the experiments.

2.1 Dataset

The dataset used in the experiments consists of parallel corpora in the language pairs EN-EL and EN-IT. The dataset is part of the test sets developed in the TraMOOC project [12]. They are educational corpora from video lectures and they contain mathematical expressions, URLs and many special characters, such as /,@,#. The corpora are described in detail by [15], [17]. The EN-EL corpora consists of 2686 segments and the EN-IT consist of 2745 segments. Two MT outputs were used - one generated by the Moses SMT toolkit [11] and the other generated by the NMT Nematus toolkit [22]. Both models trained on in- and out-of-domain data. In- and out-of-domain data included widely known corpora e.g. TED, OPUS. In order to improve the classification, a professional translation is provided for every segment. More details on the training datasets can be found in [27].

2.2 The feature set used

The feature set used is based on linguistics features divided in three categories: i) string similarity features, such as ratios between words of *S1*, *S2* and *Sr*, word distances (e.g. Dice distance [20]), percentage of segments similarity, ii) features finding the percentage of the noise in the data set (e.g. repeated words) and iii) features using length factor (LF) [21]. More details on the feature set used can be found in [17]. In this work, in order to check if the information from *SSE* will help the accuracy, additional features from the *SSE* in the EN language are used. Based on the other features, it is observed that features containing ratios are more effective to the classifier. These features are: 1) the words and character length of the *SSE*, 2) the ratio between these lengths in the *SSE* and the two MT outputs, 3) the longest word length, 4) the ratio between longest words from *SSE* and the two MT outputs and *Sr* translation.

2.3 Word Embeddings

The use of word embeddings helped us to model the relations between the two translations and the reference. In these experiments, hand-crafted embeddings were used, for the two MT outputs and the reference translation as well for both language pairs. The encoding function used is the one-hot function. The size, in number of nodes, of the embedding layer is 64 for both languages. The input dimensions of the embedding layers are in agreement with the vocabulary of each language (taking into account the most frequent words): 400 for the EN-EL language pair and 200 for the EN-IT language pair. The embedding layer used is the one provided by Keras [10] with TensorFlow as backend [1].

2.4 The NLP metrics used

The NLP set used in these experiments contains the BLEU score, METEOR, TER and WER. To calculate the BLEU score, an implementation of the BLEU score from the Python Natural Language Toolkit library [13] is used. For the calculation of the other three metrics, the code from GitHub [6] is used. All metrics were calculated for (*S1*,*S2*), (*S1*,*Sr*), (*S2*,*Sr*).

2.5 The DL schema

This study approaches the MT evaluation problem as a classification task. In particular, two volunteer linguists-annotators chose the better MT output. The linguists annotate the corpora as follows: *Y=0* if *S1* is better than *S2*, and *Y=1* if *S2* is better than *S1* for both language pairs. Where *Y* is the output, i.e. the label of the classification class. This information is used as the 'ground truth'. As an input to the learning schema, the vectors (*S1*, *S2*, *Sr*) were used, in a parallel setting. The embedding layer (as described in section 2.3) is applied and the respective embeddings *EmbS1*, *EmbS2* and *EmbSr* were created. The embeddings *EmbS1*, *EmbS2* and *EmbSr* were contracted in a pairwise setting, and the vectors (*EmbS1*, *EmbS1*), (*EmbS1*, *EmbSr*) and (*EmbS2*, *EmbSr*) were created. These vectors are the input to the hidden layers *h12*, *h1r*, *h2r* respectively. Using hidden layers *h1r* and *h2r*, the similarity between the two MT outputs and the professional translation (*Sr*) is explored. It is important to investigate the similarity between *S1* and *S2*, so an extra hidden layer *h12* is added. Interestingly, it is often observed that the MT outputs were more similar to each other than to the Sr. Every hidden layer *h12*, *h1r*, *h2r*, got as an extra input 2D matrixes $H_{12}[i,j]$, $H_{1r}[i,j]$, $H_{2r}[i,j]$, where *i* is the number of segments and *j* is the number of features. These matrices contain information about (i) the NLP set for *S1-S2*, *S1-Sr*, *S2-Sr* (as described in section 2.4) or (ii) information about linguistic features of the *SSE*, i.e. n-grams, or (iii) the combination of the previous two options. The outputs of the hidden layers *h12*, *h1r*, *h2r* are grouped and became the input to the last layer of the NN model. An extra 2D *A[i,j]* matrix with hand-crafted features (string-based) (as described in section 2.2) was added to this last layer.
The model of the DL architecture is shown in Fig. 1.

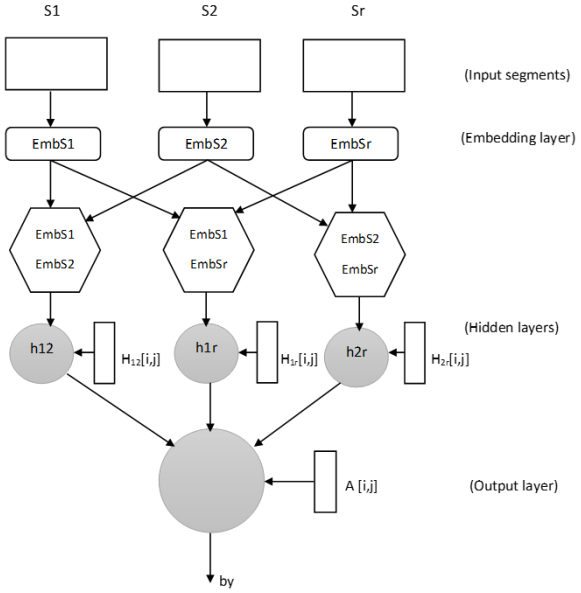

Fig. 1. Proposed model architecture

A suitable function to describe the input-output relationship in the training data should be selected. The output label is modeled as a random variable in order to minimize the discrepancy between the predicted and the true labels – maximum likelihood estimation. The binary classification problem is modeled as a Bernoulli distribution (eq. 1)

$$Y\sim \text{Bernoulli}(Y/b_y)) \tag{1}$$

Where b_y is the sigmoid function $\sigma(w^Tx+b)$, w^T and b are network's parameters.
Finally, the MaxAbsScaler [19] is used, as a preprocessing method for *EmbS1*, *EmbS2*, *EmbSr* and matrices $H_{12}[i,j]$, $H_{1r}[i,j]$, $H_{2r}[i,j]$, $A[i,j]$ as well. Every feature is scaled by its maximum absolute value.

3 Experimental Setup and Results

This section describes the details about experiments and its results.

3.1 Network Parameters

After experimentation, in order to test the proposed DL architecture, the model architecture for the experiments is defined as follows (Table 1).

Table 1. Model parameters

	Proposed NN	+NLP	+SSE	+NLP+SSE
Number of LSTM layers / Hidden Units	2/100	2/400	2/800	2/400
Dropout of LSTM Layers	0.2	0.7	0.7	0.7
Size of Dense layers / Hidden Units	3/50	3/50, 1/400	3/50, 1/800	3/50, 1/400
Dropout of Dense 4 Layer	0.2	0.7	0.7	0.7
Activation Function of Dense Layers	Relu	Relu, Linear	Relu, Linear	Relu, Linear
Output layer	Activation Sigmoid			
Learning rate	0.01	0.01	0.01	0.05
Activation Function of Dense Layers	Softmax			
Loss function	Binary cross entropy			
Optimizer	Adam (Kingma and Lei Ba, 2014)			
Batch size	256	128	64	64
Epochs	10	10	6	20

3.2 Evaluation scores

There are many machine learning evaluation metrics. In this study, commonly used metrics in classification (precision, recall and F-score) were used for the model performance evaluation. The first score (precision) shows the number of the correctly predictive values, the second score (recall) shows the percentage of total results correctly classified by the model. However, because of the unbalanced precision and recall, F-score (F1), which is a harmonic mean of precision and recall, is used. It is important to analyze the relationship between the MT outputs and the human translation, using a statistic metric - Kendall τ [9]. It is a non-parametric test used to measure the ordinal association between the two MT outputs. Kendall τ is calculated for every language pair and the macro average across all language pairs.

3.3 Results

The main results of the experiments are shown in Table 2. Different experiments were tested in the same DL architecture - using different information. The NLP set gave 67% accuracy for EN-EL and 60% for EN-IT. Subsequently, the goal was to verify if the *SSE* information can improve the model accuracy. Indeed, an increase of 2% of the classification accuracy for EN-EL and EN-IT is observed. Better accuracy results are reported when the proposed NN model uses both the information from the NLP set and *SSE* (72% accuracy for EN-EL / 70% for EN-IT). It's quite interesting that when the proposed NN model is used, without using any extra information in the hidden layers, it correctly classifies all the instances for the NMT class. Nevertheless, this model cannot be considered as the best, because the number of the correctly classified instances for the SMT class was low. The 2D matrixes *H12[i,j]*, *H1r[i,j]*,

H2r[i,j] utilization in every hidden layer *h12,h1r, h2r* gave balance between the correct instances.

Table 2. Accuracy percentage for SMT and NMT for both languages pairs.

Model	Precision		Recall		Precision		Recall	
Language pair	EN-EL				EN-IT			
	SMT	NMT	SMT	NMT	SMT	NMT	SMT	NMT
Proposed NN + NLP set	58%	69%	40%	83%	50%	65%	40%	80%
Proposed NN + SSE	69%	74%	44%	90%	55%	68%	42%	84%
Proposed NN + NLP set + SSE	68%	75%	50%	92%	62%	70%	44%	87%

It is important to have balance of accuracy performance for both classes, so the F1 score is used. In order to make a direct comparison with other models [3, 23], additional experiments were run, using, for some of them, the WEKA framework as backend [22] for the SVM and RF classifiers. It is observed that the proposed model achieves a better F1 score 4% compared with the RF, and 5% with SVM (Fig 2).

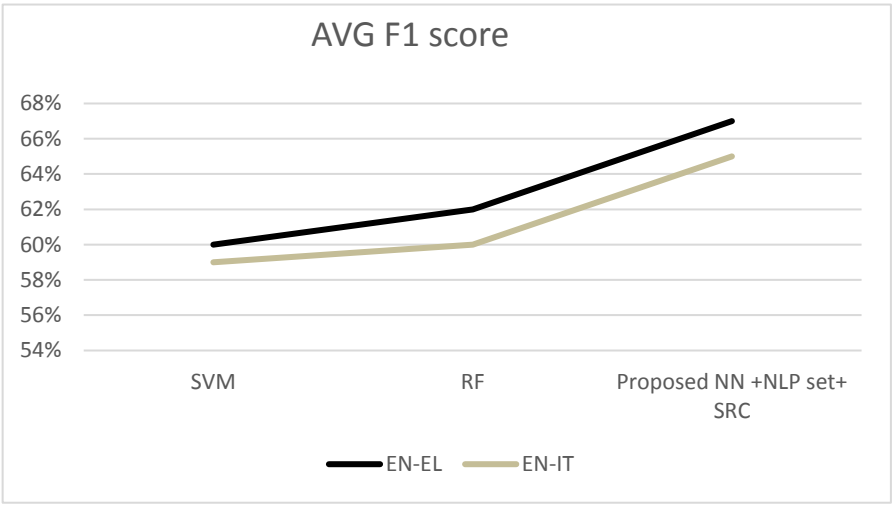

Fig. 2. Average F1 comparison between the proposed model and other works.

Table 3 shows the Kendall τ results for different models. Firstly, Kendall τ is presented for four commonly used metrics in MT evaluation (NLP set), comparing the MT outputs *S1*, *S1* with the reference *Sr*. These metrics achieved Kendall τ between 14-20. However, when they were used as extra input to the hidden layers, they led to significant improvements. In Table 2, Kendall τ values are presented for the model using different configuration setups. The NN network itself achieves lower τ value

compared to the other NN architectures, something which should not be surprising because this architecture does not use any further linguistic information. The NLP set utilization in the NN gets Kendall *τ* average (AVG) for both languages 27 points. This is because NLP metrics contain significant linguistic information about the languages (i.e. similarity scores, length). An increase up to 2.5 points is observed using information about the *SSE* (in English). Moreover, the Kendall *τ* reaches its highest value when both the NLP set and *SSE* information were applied (36 for EL / 32 for IT).

Table 3. Kendall τ for every language pair and their average.

System	EL	IT	AVG
NLP metrics set			
BLEU	17	14	15.5
METEOR	20	18	19
WER	18	16	17
TER	19	17	18
DL architecture			
Proposed NN + NLP set	29	25	27
Proposed NN +SSE	31	28	29.5
Proposed NN + NLP set +NLP set + SSE	36	32	34

3.4 Linguistic analysis

Linguistic analysis helps us to understand better the reasons why the MT output that belongs to NMT class yields higher accuracy and Kendall *τ* scores in both languages pairs. In Table 4, two cases are presented in the EN-EL language pair that the model didn't classify correctly.

Table 4. Examples of EN-EL segments.

ID	SSE	S1	S2	Sr
1	Decisions are often taken by habit, by bandwagon (everybody's doing it, so it must be right), by gut feeling.	Οι αποφάσεις λαμβάνονται συχνά από συνήθεια, με ρεύμα (όλοι το κάνουν, οπότε πρέπει να είναι σωστό), από ένστικτο.	Οι αποφάσεις συχνά λαμβάνονται από τη συνήθεια, με την άμαξα (όλοι το κάνουν, άρα πρέπει να είναι σωστό), με το ένστικτο του εντέρου.	Οι αποφάσεις παίρνονται συνήθως λόγω συνήθειας, λόγω μαζικής τάσης (όλοι το κάνουν, άρα πρέπει να είναι σωστό), λόγω καλού προαισθήματος.
2	According to Robert Pratten, what is the difference between	Σύμφωνα με τον Robert Pratten, ποια είναι η διαφορά μεταξύ trans-	Σύμφωνα με τον Ρόμπερτ Πράτεν, ποια είναι η διαφο-	Σύμφωνα με τον Robert Pratten, ποια είναι η δια-

franchise transmedia and portmanteau transmedia?	media franchise και σύμμειξη transmedia;	ρά μεταξύ των τρανζίστορ και των τρανζίσον;	φορά μεταξύ μεθοδολογίας franchise transmedia και μεθοδολογίας portmanteau transmedia;

ID 1:

- In this segment, *S2* made two serious mistakes. In the literal sense, the compound word *bandwagon* is a *wagon used for carrying a band in a parade or procession*. As a metaphor, the word *bandwagon* is used for *an activity, cause, that is currently fashionable or popular and attracting increasing support*. *S2* "didn't know" the metaphorical meaning of the word, so it has erroneously translated only the second part of the compound word in question: *wagon as άμαξα* (carriage, coach). Moreover, it is surprising that *S2* didn't even translate the first part of that compound word (*band*).
- *S2* has the phrase *gut feeling*. Gut feeling is an idiom, meaning *an instinct or intuition, an immediate or basic feeling or reaction without a logical rationale*. *S2* has literally translated the phrase: *το ένστικτο του εντέρου* (!) (the instinct of the gut). Even though in English there is also the idiom *gut instinct*, as a synonym of *gut feeling*, in Greek the literal translation of *gut instinct* is non-sensical.
- Finally, *S2* also made a slight mistake. It erroneously translated the adverb phrase *by habit* (habitually) literally: *από τη συνήθεια* (from the habit).
- *S1* has erroneously translated the above adverbial phrase *by bandwagon* as *με ρεύμα*, being unclear as to the precise meaning of the word *ρεύμα*, as in Greek this is a polysemous term that may refer to: electricity, drift, current, stream. With the preposition *με*, the Greek version is closer to the first meaning: with electricity (!), but this is nonsensical.

ID2:

- *S1* has not localized the proper noun *Robert Pratten* and rightly so, as this is the most common choice.
- *S1* did not at all translate the first of the two phrases: *francise transmedia* as well as the second word of the second phrase: *portmanteau transmedia*. *S1* has only translated the first word of this phrase: *portmanteau*, without, nevertheless, adopting the very common sense of the word: *bag, luggage, valise*, but a special and relatively rare one: *σύμμειξη* (*compounding, blending*). The professional linguist did not at all translate these phrases.
- On the contrary, *S2* translated the same phrases in a completely erroneous way: *τρανζίστορ* (*transistor*) and *τρανζίσον* (no meaning in Greek) respectively. *S2* translated these phrases incompletely and erroneously, obviously "misled" by the prefix: –*trans* of *transmedia.*

- Neither *S1* nor *S2* identified that *franchise transmedia* and *portmanteau transmedia* are methodologies (methods, techniques, approaches), as professional linguist (Reference) did.

4 Conclusion and Future Work

In this study, it is presented a DL architecture for classifying the best MT output between two options provided (one from an SMT model and the other from an NMT model), given a reference translation and an annotation schema, as well. It is worth mentioning that the translation was from EN to EL, and EN to IT which increased the task complexity, since the Greek and Italian languages are both morphologically rich languages. Well known NLP metrics were calculated and became extra inputs to the NN. Also, linguistics features from the *SSE* were used. The model's accuracy performance was tested in configurations. When the NN network combines embeddings, the NLP set (BLEU, METEOR, TER, WER) and *SSE* information (i.e. some ratios) achieved better accuracy results (increase up to 5%) and a higher Kendall τ score (increase up to 4 points) compared to related work. A linguistic analysis is also provided in order to explain linguistically the above results.
In future work, it is important to study other aspects which are likely to improve the DL architecture accuracy, such as a) a different NN configuration (e.g. different kinds of NN layers, batch normalization, learning rate), b) a feature selection method to reject the features that aren't effective for the model and c) a feature importance method to apply the proper feature weights during the NN training. In addition, it worth exploring the reasons for which the proposed model presents low accuracy values in the EN-IT pair, even though it is language independent. Finally, the model will be tested with another dataset, including in- and out-of-domain data.

Acknowledgments. This project has received funding from the GSRT for the European Union's Horizon 2020 research and innovation program under grant agreement No 644333.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265-283. USENIX Association, USA (2016).
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of 3th International Conference on Learning Representations, pp. 1–15. ICLR, San Diego (2015).
3. Barrón-Cedeño, A., Màrquez Villodre, L., Henríquez Quintana, C.A., Formiga Fanals, L., Romero Merino, E., May, J.: Identifying useful human correction feedback from an on-line machine translation service. In: Proceedings of 23rd International Joint Conference on Artificial Intelligence, pp. 2057–2063. AAAI Press, Beijing (2013).

4. Denkowski, M., & Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the 9th workshop on statistical machine translation, pp. 376-380. ACL, Baltimore, Maryland, USA (2014).
5. Duh, K.: Ranking vs. regression in machine translation evaluation. In: Proceedings of the 3rd Workshop on Statistical Machine Translation, pp. 191-194. ACL, Columbus, Ohio (2008).
6. GitHub, https://github.com/gcunhase/NLPMetrics, last accessed 2020/2/20.
7. Guzmán, F., Joty, S., Màrquez, L., & Nakov, P.: Pairwise neural machine translation evaluation. arXiv preprint arXiv:1912.03135. In: Proceedings of the 53rdAnnual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing pp. 805-814. ACL, Beijing, China (2015).
8. Jaitly, N., Sussillo, D., Le, Q. V., Vinyals, O., Sutskever, I., & Bengio, S.: A neural transducer. Cornell University Library arXiv preprint *arXiv preprint arXiv:1511.04868*. (2015).
9. Kendall, M.: A new Measure of Rank Correlation. Biometrika 30(1/2), 81–93 (1938).
10. Keras: Deep learning library for theano and tensorflow, https://keras. io/k 7.8, last accessed 2020/2/20.
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Dyer, C.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. ACL, Prague (2007).
12. Kordoni, V., Birch, L., Buliga, I., Cholakov, K., Egg, M., Gaspari, F., Georgakopoulou, Y.,Gialama, M., Hendrickx, I.H.E., Jermol, M., Kermanidis, K., Moorkens, J., Orlic, D.,Papadopoulos, M., Popovic, M., Sennrich, R., Sosoni, V., Tsoumakos, D., Van den Bosch,A., van Zaanen, M.; Way, A.: TraMOOC (Translation for Massive Open Online Courses):providing reliable MT for MOOCs. In: Proceedings of the 19th Annual Conference of theEuropean Association for Machine Translation (EAMT), pp. 376–400. European Association for Machine Translation (EAMT), Riga, (2016).
13. Loper, E., & Bird, S.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pp. 63-70. ACL, USA (2002).
14. Luong, M. T., Pham, H., & Manning, C. D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421. ACL, Lisbon, Portugal (2015).
15. 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, Cham (2018).
16. Mouratidis, D., & Kermanidis, K. L.: Comparing a Hand-crafted to an Automatically Generated Feature Set for Deep Learning: Pairwise Translation Evaluation. In: 2nd Workshop on Human-Informed Translation and Interpreting Technology, p. 66-74. HiT-IT, Varna, Bulgaria (2019).
17. Mouratidis, D., & Kermanidis, K. L.: Ensemble and Deep Learning for Language-Independent Automatic Selection of Parallel Data. Algorithms, 12(1), 12-26 (2019).
18. 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 on association for computational linguistics, pp. 311-318. Association for Computational Linguistics, Philadelphia, (2002).

19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J.: Scikit-learn: Machine learning in Python. Journal of machine learning research 12, 2825–2830 (2011).
20. Peris, Á., Cebrián, L., Casacuberta, F.: Online Learning for Neural Machine Translation Post-editing. Cornell University Library arXiv preprint 1, pp. 1–12. arXiv:1706.03196 (2017).
21. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic identification of document translations in large multilingual document collections. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), pp. 401–408. Recent Advances in Natural Language Processing (RANLP), Borovets (2003).
22. Sennrich, R., Firat, O., Cho, K., Birch-Mayne, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A., Mokry, J., Nădejde, M.: Nematus: a toolkit for neural machine translation. In: Proceedings of the EACL 2017 Software Demonstrations, pp. 65–68. ACL, Valencia (2017)
23. Singhal, S., Jena, M.: A study on WEKA tool for data preprocessing, classification and clustering. Int. J Innovative Technol. Explor. Eng. (IJITEE) 2(6), 250–253 (2013).
24. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231, Cambridge, The Association for Machine Translation in the Americas (2006).
25. Su, K. Y., Wu, M. W., & Chang, J. S.: A new quantitative quality measure for machine translation systems. In: Proceedings of the 14th conference on Computational linguistics, Volume 2 pp. 433-439. Association for Computational Linguistics, Nantes, France (1992).
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. In: 31st Conference on Neural Information Processing Systems, pp. 5998-6008. NIPS, Long Beach, CA, USA (2017).
27. Vilelmini Sosoni, Katia Lida Kermanidis, Maria Stasimioti,Thanasis Naskos, Eirini Takoulidou, Men-73 no van Zaanen, Sheila Castilho, Panayota Georgakopoulou, Valia Kordoni, and Markus Egg.: Translation Crowdsourcing: Creating a Multilingual Corpus of Online Educational Content. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, pp.479-483. European Language Resources Association, Japan (2018).
28. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. (2016).