Italian-Chinese Neural Machine Translation: results and lessons learnt

Italian-Chinese Neural Machine Translation: results and lessons learnt

Giovanni Delnevo[1*], Marcus Im[2,3], Rita Tse[2,3], Chan-Tong Lam[2,3], Su-Kit Tang[2,3], Paola Salomoni[1], Giovanni Pau[1], Vittorio Ghini[1], Silvia Mirri[1]

[1*]Department of Computer Science and Engineering, University of Bologna, Via Dell'Università 50, Cesena, 47521, FC, Italy.
[2]Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao SAR, China.
[3]Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao SAR, China.

*Corresponding author(s). E-mail(s): giovanni.delnevo2@unibo.it;
Contributing authors: marcusim@ipm.edu.mo; ritatse@ipm.edu.mo; ctlam@ipm.edu.mo; sktang@ipm.edu.mo; paola.salomoni@unibo.it; giovanni.pau@unibo.it; vittorio.ghini@unibo.it; silvia.mirri@unibo.it;

Abstract

Today, access to the Internet provides access to various forms of knowledge like free online lecture series offered by prestigious universities, massive open online courses, films and books, and Wikipedia. In addition, it is possible to join online communities on any topic of interest, get to know people with common interests, exchange thoughts and participate in debates. To enable access to these unprecedented knowledge bases, it is crucial to be able to translate texts into any language known by users. For this reason, Machine Translation has been a very active research field for the last thirty years.

In this paper, we investigate the task of Chinese-Italian translations by exploiting Neural Machine Translation approaches. We trained several deep neural networks starting from two already available datasets containing Chinese-Italian parallel corpora. Then, we compared their performance against some of the most common machine translation services freely available online. In particular, we take advantage of Microsoft Translator, Google Translate, DeepL, and ModernMT.

Keywords: Neural Machine Translation, Italian Chinese Translation, Transformer, Big Data

1 Introduction

From the seminal work of Warren Weaver [1], published in 1949, the field of automatic Machine Translation (MT), has been a very active and attractive field of study. Over the years, several approaches have been developed with the aim of providing high-quality translations automatically [2]. A first family of approaches falls under the umbrella of Rule-based MT systems [3]. It consists of learning and storing linguistic knowledge of the source and target languages: syntactic and semantic information of the words, their context, and world knowledge as a set of relationships and rules. The second type of approach is Statistical Machine Translation (SMT) [4]. These techniques compare two bilingual parallel corpora, one written in the source language and one in the target language, to find patterns and relationships in the corpora with the aim of building a statistical model that employs such patterns to transform a text in the source language into the target language. SMT techniques can be further divided into translation based on words, translation based on phrases, and translation based on syntax. Subsequently, Hybrid approaches have been investigated with the aim of combining rule-based approaches with other techniques [5]. Finally, there are Neural Machine Translation (NMT) approaches [6] that have been a major development in the field of MT. Thanks to the modern resurrection of neural networks and in particular deep neural networks, they have also been applied to MT with the aim of better capturing the relationships between the texts of two or more languages.

The rise of NMT has led to the development of many translation services, such as Google Translate and Microsoft Translator. These services continuously improve the accuracy of their translations and the languages supported. Just to cite an important happening regarding MT in 2022, there is the release of NLLB-200 by Meta[1]. NLLB stands for No Language Left Behind and it is the Meta AI model for MT. In particular, they released a single model able to translate 200 different languages, with a particular focus on some African and Indian-based languages, with the aim of increasing the number of people who can access online content or participate in digital conversations in their native languages. The release of the model was accompanied by the FLORES-200 dataset, which allows for the evaluation of translations in 40,000 different language directions. On this dataset, NLLB-200 scored an average of 44% higher performance compared to previous AI research.

In this paper, we investigate NMT approaches for Italian-Chinese translations. We aim to understand if training a custom deep neural network with the Italian-Chinese parallel corpora currently available allows to get a domain-specific model with better performance than common translation services freely available online. We employed two different datasets. One was extracted from news articles from an Italian newspaper

[1] https://about.fb.com/news/2022/07/new-meta-ai-model-translates-200-languages-making-technology-more-accessible/

which provides also the corresponding Chinese translation [7] while the other consists of parallel corpora extracted from the subtitles of more than 1,100 TEDx talks [8]. We conducted several experiments training many deep neural networks in different ways. Then, we evaluated the accuracy of the translations on the test set obtained using some of the most common MT services available online, also employing some domain adaptation mechanisms provided by such services. This allowed us to obtain a baseline for comparison, not only for this work but also for future attempts. Finally, we also exploit the domain adaptation mechanisms made available by ModernMT to evaluate if providing the parallel corpora that compose the training set allows to improve the accuracy of the translations.

The remainder of the paper is structured as follows. Section "Background and Related Work" illustrates some studies related to our research. Then, Section "Materials and Methods" describes our approach, detailing the dataset used, the methodology and the algorithms employed, and the evaluation metric. Section "Results and Discussion" illustrates the results obtained with the consequent discussion. Finally, the last Section concludes the paper with some final remarks and future works.

2 Background and Related Work

For a long time, the MT task was mainly formalized through statistical techniques, named Statistical Machine Translation (SMT) [9]. Such approaches consisted of the scrupulous crafting of features with the aim of extracting implicit information from bilingual corpora with pairs of aligned sentences. The fact that the features were not automatically extracted from the sentences but were manually defined was a crucial aspect of SMT approaches but also an important limitation [10].

The field has been revolutionized by the advent of Neural Machine Translation (NMT) systems [11, 12], becoming the dominant paradigm in MT research [13]. NMT approaches have several advantages, including the automatic extraction of features with none or very little feature engineering [14] has witnessed by the overall increase of the quality of translations observed in the literature. The rapid and significant improvements obtained by NMT systems have increased the attention on MT from both the research community and the industry, as reflected by the explosion of scientific publications related to NMT in the past few years and by the numerous NMT toolkits developed [15].

In fact, several NMT toolkits were released on Github over the years [16], including: Fairseq, Marian, ModernMT, Nematus, Neural Monkey, NiuTrans.NMT, NMT-Keras, OpenNMT, Sockeye, and THUMT. Among these toolkits, we decided to employ OpenNMT, which is written in Python employing Keras and has approximately 7k stars on GitHub.

OpenNMT [17] was designed focusing on system efficiency, code modularity, and model extensibility. We installed OpenNTM on a docker image with Tensorflow GPU already configured and we exploited its command line interface to handle the preprocessing and the training phases. It provides several ready-to-use models. Some of them are state-of-the-art models such as ListenAttendSpell [18], LuongAttention [19],

Table 1 Open-source NMT tools available on GitHub. The number of stars is relative to July 2022.

Name	Language	Framework	GitHub Stars (k)
Fairseq	Python	PyTorch	18.5
Marian	C++	-	0.9
ModernMT	Python/Java	PyTorch	0.3
Nematus	Python	Tensorflow	0.8
Neural Monkey	Python	TensorFlow	0.4
NiuTrans.NMT	C++	-	0.1
NMT-Keras	Python	Keras	0.5
OpenNMT	Python/C++	PyTorch/TensorFlow	7
Sockeye	Python	PyTorch	1.1
THUMT	Python	PyTorch/TensorFlow/Theano	0.6

LstmCnnCrfTagger [20], LstmCnnCrfTagger [20], and TransformerBaseSharedEmbeddings [21]. Some of them have also tiny and big versions such as TransformerTiny and TransformerBig. Some general bidirectional LSTM encoder-decoder models are available, with different sizes (e.g., NMTSmallV1, NMTMediumV1, and NMTBigV1). Finally, it also provides a small version of GPT-2 [22].

3 Materials and Methods

This Section illustrates the datasets employed in the study, the adopted methodology together with the approaches for neural machine translation, and the evaluation metric used to evaluate the translations.

3.1 Dataset Description

In this context, it is crucial the definition of suitable and adequate parallel corpora [23–25]. This is why, in our experiments, two different Italian-Chinese parallel corpora were used. The former corpus has been extracted from news articles from an Italian newspaper which provides also the corresponding Chinese translation [7].

The latter is MulTed [8], which consists of parallel corpora extracted from the subtitles of more than 1,100 TEDx talks. The dataset has been constructed starting from the TED talks library, which contains the recordings of independently non-profit organized events in over 130 countries. 102 languages are available in the datasets, from the most common such as English, Spanish, and other European languages, including Italian, passing through languages of African countries such as Afrikaans and Swahili, up to languages of Asian countries such as Japanese, Chinese (in both traditional and simplified versions) and Thai. They are bilingually aligned at the sentence level using English as the pivot language.

In the end, we were able to obtain 352,449 sentences that were divided into training (276,959), validation (5,000), and test (70,490) sets. From now on, we will refer to such a dataset as the *Talk Dataset*.

3.2 Methodology and Approaches

As anticipated in the Introduction Section, in this work, we want to investigate if a direct translation can be carried out without the intermediate translation in the English language. We conducted several experiments directly training a deep neural network using the two datasets, the News and the Talks ones. In the experiments, we considered:

- Both directions of translation, from Italian to Chinese and from Chinese to Italian.
- The News and the Talks datasets alone or combined, by training the deep neural network on one dataset and evaluating its performance on the other one.

In the experiments, we employed the Transformer architecture proposed in [21], composed of six encoders and six decoders. We also evaluated a tiny version of the Transformer with just two encoders and two decoders. Both models were implemented using OpenNMT [17].

For the Neural Machine Translation Services, instead, we considered the following ones. **Google Translate** [26]. It is the production NMT system at Google. Initially, it was designed to overcome three inherent weaknesses of NMT: its slower training and inference speed, ineffectiveness in dealing with rare words, and sometimes failure to translate all words in the source sentence [27]. It currently supports the translation of more than 1,000 languages. It supports "Zero-Shot Translation" which is a translation between language pairs never seen explicitly by the system [28]. **Microsoft Translator** [29]. It is the translation system developed by Microsoft and it is also based on NMT. **ModernMT** [30]. It has been developed within a three-year Horizon 2020 innovation action that had the aim of developing new open-source machine translation technology for use in translation production environments, both fully automatic and as a back-end in interactive post-editing scenarios. Its main functionalities include simple installation procedures, fast setup times for systems built from scratch, immediate integration of new data, instant domain adaptation, and high scalability. **DeepL** [31]. It is a platform launched in 2017 by a German company. According to several comparison studies, it outperforms the other translation systems, even if it supports a limited number of languages (i.e., 22) compared to the other translation systems.

3.3 Evaluation Metric

As far as the evaluation metric is concerned, the translation system is evaluated with the BiLingual Evaluation Understudy (BLEU) score [32]. It is a corpus-based metric. The idea is to look for exact matches of words, considering the word precision and without considering the recall. The final score is computed as the weighted geometric average of the n-gram scores computed for n from 1 to 4.

The brevity penalty is introduced since the recall is not considered and aims to penalize translations that are too short compared to the closest reference length with exponential decay. The N-Gram Overlap, instead, counts how many unigrams, bigrams, trigrams, and four-grams match their respective n-gram counterparts.

According to the Google Translate documentation[2], the BLEU score can be interpreted according to Table 2. BLEU scores higher than 30 are considered understandable while scores higher than 50 are considered high quality.

Table 2 Bleu Score interpretation provided by the Google Translate documentation.

BLEU Score	Interpretation
<10	Almost useless
10 -19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
>60	Quality often better than human

4 Results and Discussion

This Section presents the results obtained in the three phases of our methodology: i) with the Transformer model, ii) with the Translation Services, and iii) with the domain adaptation mechanism of ModernMT.

4.1 Neural Machine Translation

We began the experiments with a cross-dataset evaluation. The idea is to train the model with the training data of a dataset and to test it using the testing data of the other dataset. After building a common vocabulary for the two datasets, we trained the models considering both directions of translation (i.e., Italian to Chinese and Chinese to Italian). We employed a simple YAML configuration file, where we specified just the directories of the data. Once the training was completed, we evaluated the performance of the model using the testing set of the other dataset (i.e., the one not used for training).

The results of the cross-dataset evaluation are reported in Table 3. As shown, the model is not able to translate sentences in both directions, as highlighted by the low BLEU score that never surpasses one. We hypothesized that the fact that the two datasets contain text relative to different domains (i.e., the newspaper articles and the transcription of talks) led to an even more difficult task. Furthermore, the News dataset contained far few examples than the Talks dataset. Anyway, this justifies only the bad results obtained when training the model on the News dataset and testing it on the Talks one.

Hence, we decided to evaluate the two datasets separately. We built two vocabularies, one for each dataset and we launched the training just specifying in the YAML configuration file the paths of the files. Anyway, the situation did not improve, as witnessed by the results reported in Table 4. While the result using the News dataset was

[2] https://cloud.google.com/translate/automl/docs/evaluate

Table 3 BLEU Score obtained using the Transformer on cross-dataset evaluation

Train Dataset	Test Dataset	Translation Direction	Bleu Score
Talks	News	it ->ch	<1
Talks	News	ch ->it	<1
News	Talks	it ->ch	<1
News	Talks	ch ->it	<1

somehow expected due to the limited amount of examples in the dataset, the result on the Talks dataset was not. In both cases, the BLEU score computed on the test set was less than one.

Table 4 BLEU Score obtained using the Transformer on the two dataset separately.

Train Dataset	Test Dataset	Translation Direction	Bleu Score
Talks	Talks	it ->ch	<1
Talks	Talks	ch ->it	<1
News	News	it ->ch	<1
News	News	ch ->it	<1

As the final experiments, we decided to evaluate the two datasets together. We re-used the vocabulary created for the first set of experiments and we used the same simple configuration file. We did not expect particular improvements with respect to the previous experiments and the results, reported in Table 5, met our expectations, with the BLEU score on the test set always being less than one.

Table 5 BLEU Score obtained using the Transformer on the two dataset merged.

Train Dataset	Test Dataset	Translation Direction	BLEU Score
Talks + News	Talks + News	it ->ch	<1
Talks + News	Talks + News	ch ->it	<1

We hypothesized that the bad results obtained in the first set of experiments could be due on the one hand to a large number of parameters to be optimized and on the one hand to the scarce quantity of examples in the dataset. For this reason, we repeated the same experiments varying the model used. Instead of a Transformer that has six encoders and six decoders, we employed a tiny version of it, made available by OpenNMT which is the TinyTransformer. Such a model has just two encoders and two decoders. The BLEU scores obtained in the different experiments using the TinyTransformer are reported in Table 6. As shown, there are no differences with the ones obtained using the Transformer.

According to these results, the high number of parameters was not an issue. Rather it is the datasets employed in the experiments. The reasons behind the bad results are manifold. First, the languages chosen for this case study, Italian and Chinese, are pretty different. While Italian is a Latin language, Chinese is not. Furthermore, in the

Table 6 BLEU Score obtained in the various experiments using the TinyTransformer.

Train Dataset	Test Dataset	Translation Direction	BLEU Score
Talks	News	it ->ch	<1
Talks	News	ch ->it	<1
News	Talks	it ->ch	<1
News	Talks	ch ->it	<1
Talks	Talks	it ->ch	<1
Talks	Talks	ch ->it	<1
News	News	it ->ch	<1
News	News	ch ->it	<1
Talks + News	Talks + News	it ->ch	<1
Talks + News	Talks + News	ch ->it	<1

Chinese language, the words are mainly constituted by other words (or of elements with a meaning similar to words). Moreover, the characters are not a completely disconnected symbol system. They are made up of traits, which constitute a closed inventory and must be tracked in a fixed manner and in a fixed order. Most characters consist of an element that indicates the semantic category to which the represented word belongs, and another that suggests its sound. Each character is associated with a syllable, and each syllable is associated with a meaning. The number of characters should be 6,763 for simplified Chinese [33]. Another difference is that words tend to be invariable. The second issue regards the numerosity of the datasets. Unfortunately, the number of examples in the two only datasets available does not surpass 300,000 examples. Then, also the peculiarity of the nature of the data used has to be taken into consideration. In fact, the Talks dataset contains the transcription of talks made by different mother tongues people that are then translated into Italian and Chinese, with the requirement of following the rhythm of the speech. Finally, several results in the field of machine translation have shown that translation systems work better in very specific and restricted domains [34, 35].

4.2 Translation Services

Given the (poor) results obtained in the previous Subsection, we decided to evaluate the performance of some common neural machine translation services with the aim of having a baseline comparison. As already mentioned in the previous Section, we employed four different services: Google Translate, ModernMT, Microsoft Translator, and DeepL. We decided to use just a subsection of the Talks dataset testing set, randomly selecting 220 examples.

The results are reported in Table 7. As shown, there are no significant differences in the performance of all the services. All of them have better BLEU scores when translating from Chinese to Italian rather than from Italian to Chinese. DeepL obtained the best BLEU score (13.41) in the translation from Chinese to Italian while ModernMT outperformed the other services, obtaining a BLEU score of 4.96 when translating from Italian to Chinese. It is important to notice that, recalling the BLEU score interpretation presented in Table 2, the Italian to Chinese translations are "almost useless"

while the Chinese to Italian is "Hard to get the gist". Such results can be used as a baseline for future comparisons. Furthermore, they demonstrated the difficulties of the translation task on such a dataset.

Table 7 BLEU Score obtained using some common neural machine translation services.

Platform		**Azure**		**ModernMT**		**Google**		**DeepL**	
Target		**ita**	**chi**	**ita**	**chi**	**ita**	**chi**	**ita**	**chi**
BLEU		10.6	4.58	11.53	4.96	13.39	4.24	13.41	4.3
1-gram	**Individual**	33.58	18.74	37.19	18.4	38.35	17.98	35.33	12.78
2-10	**Cumulative**	33.58	18.74	37.19	18.4	37.62	17.98	35.33	12.47
2-gram	**Individual**	15.17	4.6	16.04	6.06	18.27	4.44	17.23	5.35
2-10	**Cumulative**	22.57	9.29	24.43	10.56	25.96	8.94	24.67	8.07
3-gram	**Individual**	7.52	2.5	8.2	2.54	9.88	1.82	9.85	1.87
2-10	**Cumulative**	15.65	6	16.98	6.57	18.69	5.26	18.17	4.91
4-gram	**Individual**	3.3	2.04	3.61	2.13	5.03	2.22	5.39	2.94
2-10	**Cumulative**	10.6	4.58	11.53	4.96	13.39	4.24	13.41	4.3

4.3 Domain Adaptation

In the final experiment, we evaluated if employing some domain adaptation mechanisms is possible to improve the accuracy of the translations of a translation service. All the Neural Machine Translation Services, employed in the previous set of experiments, provide some form of domain adaptation.

We chose to conduct the experiment of domain adaptation using ModernMT. ModernMT employs translation memories. Memory is an entity that stores a pair of sentences and their corresponding translation. They are employed through the context vector to perform real-time adaptation. We created translation memories employing the training set examples, after transforming it into the TMX format. Then, we re-evaluated the test set and computed the BLEU score. The results are reported in Table 8. As shown, there are no significant differences in the results. The BLEU score when translating from Chinese to Italian gets slightly worse while translating from Italian to Chinese slightly improves.

At the end of our experimentation, we can draw the following conclusions. The first thing that clearly emerges from our experiments and confirms other results in the literature is the need for a large amount of data. In the specific case of NMT, parallel corpora of sentences with their corresponding high-quality translations are required. This is not always easy to do as the labeling process is very expensive. As evidence of this, many datasets for automatic translations are based on excerpts from government documents or instruction manuals that are translated into multiple languages. Such a problem is exacerbated by the case study considered, which is focused on Italian Chinese translations. In fact, languages with structural differences and a diversity of grammar rules increase the difficulty of the problem. This is also confirmed by the results obtained using already available translation services, like Google Translate, Microsoft Translator, DeepL, and ModernMT.

Table 8 BLEU Score obtained using ModernMT, employing the domain adaptation mechanisms.

ModernMT		Standard Approach		Domain Adaptation	
Target		ita	chi	ita	chi
BLEU		11.53	4.96	11.27	5.05
1-gram	Individual	37.19	18.4	36.86	19.51
2-6	Cumulative	37.19	18.4	36.86	19.51
2-gram	Individual	16.04	6.06	15.74	6.49
2-6	Cumulative	24.43	10.56	24.09	11.26
3-gram	Individual	8.2	2.54	8.09	2.52
2-6	Cumulative	16.98	6.57	16.74	6.84
4-gram	Individual	3.61	2.13	3.44	2.04
2-6	Cumulative	11.53	4.96	11.27	5.05

5 Conclusions

This paper investigates the task of NMT, with a specific focus on the translations between Italian and Chinese. We used two Italian-Chinese parallel corpora. The former is extracted from a newspaper site that provides a translation for some articles while the other consists of the transcript of Ted Talks. We conducted several experiments, training a Transformer using OpenNMT. In none of the experiments, the algorithm is able to obtain a BLEU score greater than one.

This work paves the way for several future works. First, in order to have a baseline comparison, some translation services (such as Microsoft Translator and Google Translate) could be evaluated. Then, a high-quality and large Italian-Chinese parallel corpus could be built. This can be achieved in two ways. The former is a search for other sources that make available parallel corpora of texts in Italian and Chinese. The latter one, instead, consists in building a web application that allows to collect sentences and provides apposite interfaces for the translation [36, 37]. Moreover, different architectures of deep neural networks can be evaluated for training from scratch.

Acknowledgements. The authors are deeply grateful towards Federico Garcea for his precious suggestions. This work was supported in part by the Macao Polytechnic University (P178/FCA/2022).

References

[1] Weaver, W.: Memorandum on translation. Online (1949)

[2] Chand, S.: Empirical survey of machine translation tools. In: 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 181–185 (2016). IEEE

[3] Bahadur, P., Jain, A., Chauhan, D.: Etrans-a complete framework for english to sanskrit machine translation. In: International Journal of Advanced Computer Science and Applications (IJACSA) from International Conference and Workshop on Emerging Trends in Technology (2012)

[4] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N.A., Yarowsky, D.: Statistical machine translation. In: Final Report, JHU Summer Workshop, vol. 30 (1999)

[5] Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., Chen, Y.: Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 179–182 (2008)

[6] Koehn, P.: Neural machine translation. arXiv preprint arXiv:1709.07809 (2017)

[7] Tse, R., Mirri, S., Tang, S.-K., Pau, G., Salomoni, P.: Building an italian-chinese parallel corpus for machine translation from the web. In: Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good, pp. 265–268 (2020)

[8] Zeroual, I., Lakhouaja, A.: Multed: A multilingual aligned and tagged parallel corpus. Applied Computing and Informatics (2020)

[9] Maruf, S., Saleh, F., Haffari, G.: A survey on document-level neural machine translation: Methods and evaluation. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)

[10] Lopez, A.: Statistical machine translation. ACM Computing Surveys (CSUR) **40**(3), 1–49 (2008)

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

[12] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)

[13] Ragni, V., Nunes Vieira, L.: What has changed with neural machine translation? a critical review of human factors. Perspectives **30**(1), 137–158 (2022)

[14] Karimova, S., Simianer, P., Riezler, S.: A user-study on online adaptation of neural machine translation to human post-edits. Machine Translation **32**(4), 309–324 (2018)

[15] Stahlberg, F.: Neural machine translation: A review. Journal of Artificial Intelligence Research **69**, 343–418 (2020)

[16] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y.: Neural machine translation: A review of methods, resources, and tools. AI Open **1**, 5–21 (2020)

[17] Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810 (2017)

[18] Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell. arXiv preprint arXiv:1508.01211 (2015)

[19] Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)

[20] Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017)

[22] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[23] Cheok, S.M., Hoi, L.M., Tang, S.-K., Tse, R.: Crawling parallel data for bilingual corpus using hybrid crawling architecture. Procedia Computer Science **198**, 122–127 (2022)

[24] Cheok, S.M., Hoi, L.-M., Tang, S.-K., Tse, R.: Constructing high quality bilingual corpus using parallel data from the web. In: IoTBDS, pp. 127–132 (2022)

[25] Roccetti, M., Prandi, C., Mirri, S., Salomoni, P.: Designing human-centric software artifacts with future users: a case study. Human-centric Computing and Information Sciences **10**(1), 1–17 (2020)

[26] Bapna, A., Caswell, I., Kreutzer, J., Firat, O., Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., et al.: Building machine translation systems for the next thousand languages. arXiv preprint arXiv:2205.03983 (2022)

[27] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv (2016). https://doi.org/10.48550/ARXIV.1609.08144 . https://arxiv.org/abs/1609.08144

[28] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv (2016). https://doi.org/10.48550/ARXIV.1611.04558 . https://arxiv.org/abs/1611.04558

[29] Almahasees, Z.M.: Assessment of google and microsoft bing translation of journalistic texts. International Journal of Languages, Literature and Linguistics **4**(3),

231–235 (2018)

[30] Germann, U., Barbu, E., Bentivoglio, M., Bogoychev, N., Buck, C., Caroselli, D., Carvalho, L., Cattelan, A., Cattoni, R., Cettolo, M., Federico, M., Haddow, B., Madl, D., Mastrostefano, L., Mathur, P., Ruopp, A., Samiotou, A., Sudharshan, V., Trombetti, M., van der Meer, J.: Modern mt: A new open-source machine translation platform for the translation industry. Baltic Journal of Modern Computing **4**(2), 397–397 (2016)

[31] Rescigno, A.A., Vanmassenhove, E., Monti, J., Way, A.: A case study of natural gender phenomena in translation. a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In: CLiC-it (2020)

[32] 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)

[33] Chu, C., Nakazawa, T., Kawahara, D., Kurohashi, S.: Chinese-japanese machine translation exploiting chinese characters. ACM Transactions on Asian Language Information Processing (TALIP) **12**(4), 1–25 (2013)

[34] Wang, X., Chen, C., Xing, Z.: Domain-specific machine translation with recurrent neural network for software localization. Empirical Software Engineering **24**(6), 3514–3545 (2019)

[35] Arcan, M., Buitelaar, P.: Translating domain-specific expressions in knowledge bases with neural machine translation. CoRR, abs/1709.02184 (2017)

[36] Delnevo, G., Mirri, S., Prandi, C., Manzoni, P.: An evaluation methodology to determine the actual limitations of a tinyml-based solution. Internet of Things **22**, 100729 (2023)

[37] Salomoni, P., Mirri, S., Ferretti, S., Roccetti, M.: A multimedia broker to support accessible and mobile learning through learning objects adaptation. ACM Transactions on Internet Technology (TOIT) **8**(2), 1–23 (2008)