Machine Translation of literary texts: genres, times and systems

Ana Isabel Cespedosa
anabelcespedosa@gmail.com

Ruslan Mitkov
Lancaster University
r.mitkov@lancaster.ac.uk

Abstract

Machine Translation (MT) has taken off dramatically in recent years due to the advent of Deep Learning methods and Neural Machine Translation (NMT) has enhanced the quality of automatic translation significantly. While most work has covered the automatic translation of technical, legal and medical texts, the application of MT to literary texts and the human role in this process have been underexplored. In an effort to bridge the gap of this under-researched area, this paper presents the results of a study which seeks to evaluate the performance of three MT systems applied to two different literary genres, two novels (*1984* by George Orwell and *Pride and Prejudice* by Jane Austen) and two poems (*I Felt a Funeral in my Brain* by Emily Dickinson and *Siren Song* by Margaret Atwood) representing different literary periods and timelines. The evaluation was conducted by way of the automatic evaluation metric BLEU to objectively assess the performance that the MT system shows on each genre. The limitations of this study are also outlined.

1 Rationale

Recent advances in Artificial Intelligence and MT have brought a new perspective to the ongoing discussion on the automatic translation of literary texts among academics. More specifically, the significantly improved performance of Neural Machine Translation has triggered a debate among translation professionals about the future role of the translators.

It has been demonstrated that MT delivers better results when applied to scientific and technical texts which lack ambiguity and provide a precise message (Moorkens et al., 2018). On the other hand, literary texts are rich in rhetorical devices, ambiguity, and precise a certain level of creativity, becoming a great challenge for the MT to face when translating this type of texts (Toral, 2018), and as a result producing more literal translations which do not convey the essential meaning of the texts (Moorkens et al., 2018). Nevertheless, the research on how feasible MT for literary texts is or on the development of new techniques to improve the quality of literary MT, has been insufficient. The related work has centred on determining the main factors of the use of MT as well as identifying the cognitive effects on the human translator when applying to its workflow. These studies have proven that MT is both useful and powerful tool used in the translation process and can enhance the productivity of the human translator (Toral and Way, 2015a; Guerberof and Toral, 2020).

This study has been motivated by the recent advances of NMT and by the fact that the topic of the application of MT to literary texts has been underexplored. In particular, this study seeks for the first time to:

- identify whether MT performance is influenced by the genre and the time period of the literary texts. If so, how and to what extent do these aspects impact the MT performance.

- compare the performance of three recent NMT systems on literary texts.

The rest of the paper is structured as follows. Section 2 discusses related work. Section 3 details the methodology used by outlining the data used

Proceedings of the First NLP4TIA Workshop, pages 48–53
Varna, Sep 8, 2023
https://doi.org/10.26615/978-954-452-091-5_007

and the experiments conducted. Section 4 reports the evaluation results and provides a brief discussion of the obtained results. Section 5 lists the limitation of this study and finally Section 6 presents the conclusions.

2 Related Work

Most related work has focused on the feasibility of MT applied to scientific texts. With regard to literary MT, the studies have been focused on the narrative genre and the cognitive efforts of the human translator when MT is applied to the translation workflow.

It has been proven that the use of MT in scientific and technical texts have enhanced the human translators' productivity (Toral and Way, 2014). Since literary texts exhibit more ambiguity and use literary devices to infer its meaning, the widespread view is that MT cannot cope with this type of text (Bellos, 2012 in Toral and Way, 2015a; Kelly and Zetzsche, 2012). Nevertheless, there are researchers who consider possible the use of MT for translating literary texts successfully (Salimi, 2014, Toral and Way, 2015a; Toral and Way, 2018; Matusov, 2019).

Genzel et al. (2010) studied the Machine Translation of a poem, considering its metric, length, and rhythm. The results showed that the format could be preserved, but it could not obtain the same quality nor preserve the meaning.

Voigt and Jurafsky (2012) focused on the referential cohesion in literary and non-literary texts and their outputs when processed by MT. They concluded that although literary texts had more cohesive references than non-literary texts and although MT was able to cope with them, the referential cohesion is a key factor for good MT performance. Richardson (2012) employed Microsoft Translator Hub in the translation process of a church to cater for the demand of translation into several languages, creating corpora and glossaries and resulting in a higher productivity.

Toral and Way (2015a) reported the results of a study on the applicability of MT to literary text taking into account how related are the languages involved (French, English and Italian; Spanish and Catalan). They proposed to fine-tune the MT systems to the different types of literary texts regarding their characteristics such as cohesion, literary devices, dialogue, etc. They experimented (Toral and Way; 2018) with an NMT system customised to translate distant languages such as English and Catalan and compared the results with the previous translations by an SMT system. They drew their study on previous research (Toral and Way (2015b)) where factors such as limitation and freedom of translation were taken into consideration.

Moorkens et al. (2018) studied the perception of literary MT (based on both SMT and NMT) by human translators. Six professional translators were asked to translate from English to Catalan in three different modes: translating from scratch, post-editing NMT output and post-editing SMT output. It was shown that human translators preferred translating from scratch literary texts, but considered useful the suggestions offered by the MT systems. In line with this work, Toral and Way (2018) also proposed to incorporate automatic systems to the translation workflow to help the human translator improve their productivity.

Matusov (2019) examined the challenges that NMT faced when applied to literary texts with English and Russian as language pair and reported better performance after fine-tuning the MT systems to this particular language pair. In another study, Kuzman et al. (2019) applied NMT to Slovenian literary texts with the results showing increase in the productivity.

Omar and Gomaa (2020) identified the challenges MT systems face when translating literary texts; they concluded the most typical mistakes are pragmatic, structural and lexical. Guerberof and Toral (2020) analysed the impact of post-editing and MT on creativity and literature, as well as the perceptions of professional translators on this issue. Kenny and Winters (2020) studied how MT used in the translation process affects the translator's voice. Finally, Fonteyne et al. (2020) evaluated the recent improvements of NMT systems when applied to literary texts aiming to produce coherent translation at a textual level.

Among the most recent work is that from Ruffo (2022) who questions the lack of inclusion of literary translators in the discussion on technological advances of the automatic translation tools. A survey was conducted to identify the perception of technology applied to translation by professional translators. Although professional translators are not reluctant to technology, the negative views mostly had to do with the use and development of translation tools applied to literature.

3 Methodology: Data and experiment

Four texts from different literary genres were selected to establish their impact on the quality of MT; this study also sought to determine if the time period of the literary texts could influence the performance of MT. From the prose genre, two novels were chosen: *Pride and Prejudice* by Jane Austen, and *1984* by George Orwell; the poetry was represented by the following two poems: *I Felt a Funeral in my Brain* by Emily Dickinson, and *Siren Song* by Margaret Atwood. See Table 1 for more details. Three 3 popular NMT systems were experimented with: DeepL, Systran and Yandex.

LITERARY GENRE	ORIGINAL TEXT	TRANSLATIONS
Narrative	*Pride and Prejudice* (Jane Austen, 1813)	*Orgullo y prejuicio* (José Jordán de Urríes y Azara, 1924) *Orgullo y prejuicio* (Marta Salís, 2014)
Narrative	*1984* (George Orwell, 1949)	*1984* (Miguel Temprano García, 2013) *1984* (Rafael Vázquez Zamora, s.f.)
Lyric	*I Felt a Funeral in my Brain* (Emily Dickinson, 1858-1859)	*Sentí un Funeral, en mi Cerebro* (Álvaro Torres Ruiz, s.f.) *Sentí un Funeral, en el Cerebro* (Marta Rosillo Moya, 2021)
Lyric	*Siren Song* (Margaret Atwood, 1974)	*La canción de la sirena* (Raquel Rivas Rojas, s.f.) *El canto de la sirena* (Andrés Catalán, 2013).

Table 1. Texts selected for the study.

In order to objectively assess the quality of MT performance, we implemented the BLEU score metric (Papineni et al., 2002).

To this end, we chose two human translations from different time periods for each selected text. The BLEU metric system was set up both at sentence level with *sentence_bleu()* function and at corpus level with *corpus_bleu()* function contained in the NLTK (Natural Language Toolkit) library, as well as a cumulative score is obtained by assigning a cumulative weighting of 4-grams.

In addition to the BLEU score, the most significant linguistic features of each text were described, and the approaches taken by both the machine translation and the human translation were compared and correlated with the BLEU score.

4 Results and discussion

The evaluation results calculated by BLEU suggested that the lowest (0.3 out of 100) and highest (39.79 out of 100) quality scores were both achieved by Systran. However, reviewing the questions put forward in the Section 1 of this paper, it is safe to consider the following results.

The MT system which achieved best performance out of the three was DeepL, estimating the average BLEU score for the four texts. Overall, DeepL system was capable of producing a translation similar to the human translation with a higher consistency and quality as compared with the other systems. Although Systran and Yandex made the same type of mistakes, Yandex was more consistent and closer to DeepL (see Table 2).

	DEEPL	SYSTRAN	YANDEX
Average score for narrative genre	21.98	7.51	22.76
Average score for lyric genre	30.14	26.95	26.21
Global average score	26.06	17.23	24.48

Table 2. Global and Individual Average Score for each MT system

As for the performance of MT systems on literary genres, a higher score was achieved on poetry as opposed to prose. Systran obtained the lowest score for George Orwell's novel resulting in the decrease of the global average score for the prose. On the other hand, Margaret Atwood's poem emerged as the work with the best MT output, followed by George Orwell's novel.

Therefore, according to BLEU score, MT fared much better on poetry than prose when compared to a human translation. See Table 3.

	DeepL	Systran	Yandex
AVERAGE SCORE (Classical literature)	15.07	14.55	15.25
AVERAGE SCORE (Contemporary literature)	37.05	19.91	33.72

Table 3. Average score for each text following the results of the 3 MT systems.

Finally, all three MT systems delivered a low-quality output on the older classical works (those by Jane Austen and Emily Dickinson) as see in Table 4:

WORK	**1984**	**Pride and Prejudice**	**I Felt a Funeral, in my Brain**	**Siren Song**
AVERAGE SCORE (for 3 MT systems)	22.08	19.13	17.16	38.37

Table 4. Average score according to the temporary nature of the works.

In summary, the MT systems were able to perform better on modern literature which is expected to have a less complex style. We conjecture that another reason for that is because NMT systems are usually trained on more contemporary data.

5 Limitations

It should be noted that the results and conclusions should not be taken fully representative due to the following limitations of this study:

- **Data size**

  This type of study ideally requires larger datasets, or a large corpus in order to be significantly and sufficiently representative for the data obtained. In this study, only four texts have been used, two of them are extracts of a larger work, so the results cannot be generalised. Furthermore, the BLEU metric requires a large number of references in order for the scores to be as accurate and objective as possible. If not, there is a risk of obtaining not-so-accurate scores, since the algorithm is based on the comparison of MT and HT options.

- **BLEU limitations**

  In addition to the shortcoming mentioned in the previous paragraph regarding the number of references needed, it should be noted that this metric has its own shortcomings if not properly implemented. The algorithm does not consider the meaning of the sentence or the language variations as it regards sentences as strings. In other words, the system could compare the MT with the HT that may not be fully accurate or may contain mistakes as well. It may be the case for a good MT system to obtain a low score if it has been compared to a poor HT. Despite these shortcomings and limitations, BLEU is still one of the most widely used MT metrics.

- **Corpus representativeness**

  The representativeness of a corpus is as important as the size of the sample. In this study the literary sources were selected on the basis of their genre and availability online. For a more thorough study, it will be appropriate to choose a larger number of texts with a greater range of linguistics features in order to study to what extent the MT system can cope with these translations.

6 Conclusion

The aim of this study is to analyse the performance of the MT systems selected for

different literary genres due to the lack of literature that addresses this issue. It was sought to assess the feasibility of MT to literary texts to and revisit the generally pessimistic widespread perception questioning the use of MT within the workflow of the literary translator.

To this end, three NMT systems (DeepL, Systran and Yandex) were selected to assess the performance and quality when translating prose and poetry from different time periods. The results suggest that the best performing system on these texts according to our experiments was DeepL. This NMT system produces more coherent and similar texts to those produced by humans. In addition, the obtained BLEU scores show that: a) MT fares better on poetry and does not do so well on prose b) MT delivers better results on modern contemporary texts and does not do so well on older classic texts.

However, it is essential to acknowledge the limitations of this study as outlined in the previous section. In future studies, it will be preferrable to use larger and more representative data in the experiments. The BLEU evaluation metric could be compared and correlated with other metrics such as TER (Translation Error Rate) and WER (Word Error Rate) to.

References

Abdulfattah Omar, and Yasser Gomaa. 2020. The Machine Translation of Literature: Implications for Translation Pedagogy. In International Journal of Emerging Technologies in Learning, vol. 15(11): 228-235. https://online-journals.org/index.php/i-jet/article/view/13275/7151.

Ana Guerberof Arenas, and Antonio Toral. 2020. The Impact of Post-editing and Machine Translation on Creativity and Reading Experience. In Translation Spaces, vol. 9(2): 255-282. https://doi.org/10.48550/arXiv.2101.06125.

Antonio Toral, and Andy Way. 2014. Is machine translation ready for literature? In J. Esteves-Ferreira, J. Macan, R. Mitkov, and S. Olaf-Michael (eds), Translating and The Computer, vol. 36. https://aclanthology.org/2014.tc-1.23.pdf.

Antonio Toral, and Andy Way. 2015a. Machine-Assisted Translation of Literary Text: A Case Study. In *Translation Spaces*, vol. 4(2): 241-268. https://www.researchgate.net/publication/290209944_Machine-assisted_translation_of_literary_text_A_case_study.

Antonio Toral, and Andy Way. 2015b. Translating Literary Text between Related Languages using SMT. In A. Feldman, A. Kazantseva, S. Szpakowicz, and C. Koolen (eds.), Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pp. 123-132. https://aclanthology.org/W15-0714.pdf.

Antonio Toral, and Andy Way. 2018. What Level of Quality can Neural Machine Translation Attain on Literary Text? In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (eds.), Translation Quality Assessment (1st ed., pp. 263-287). Springer Cham. https://doi.org/10.48550/arXiv.1801.04962.

David Bellos. 2012. Is That a Fish in Your Ear?: Translation and the Meaning of Everything. London: Particular Books.

Dmitry Genzel, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. In H. Li and L. Màrquez (eds), Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 158-166. https://aclanthology.org/D10-1016.pdf.

Dorothy Kenny, and Marion Winters. 2020. Machine translation, ethics and the literary translator's voice. In Translation Spaces, vol. 9(1): 123-149. https://pure.hw.ac.uk/ws/portalfiles/portal/41374556/Kenny_Winters_Translation_Spaces_Article_accepted_24_April.pdf.

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In J. Hadley, M. Popović, H. Afli, and Andy Way (eds), The Qualities of Literary Machine Translation, pp. 19-23. https://aclanthology.org/W19-7302.pdf.

Jonni Salimi. 2014. Machine Translation of Fictional and Non-fictional Texts. An examination of Google Translate's accuracy on translation of fictional versus non-fictional texts. [Bachelor Degree Project, Stockholm University].

Joss Moorkens, Sheila Castilho, Antonio Toral, and Andy Way, A. 2018. Translators' perceptions of literary post-editing using statistical machine translation. In *Translation Spaces*, vol. 7(2): 240-262. https://www.researchgate.net/publication/329263225_Translators%27_perceptions_of_literary_post-

editing_using_statistical_and_neural_machine_translation.

Kelly Nataly, and Jost Zetzsche. 2012. Found in Translation: How Language Shapes Our Lives and Transforms the World. New York: Perigee Trade.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, and D. Lin (eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318. Association of Computational Linguistics. https://aclanthology.org/P02-1040.pdf.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (eds.), Proceedings of the 12th Conference on Language Resources and Evaluation (LREC), pp. 3783-3791. https://aclanthology.org/2020.lrec-1.468.pdf.

Paola Ruffo. 2022. Collecting literary translators' narratives: Towards a new paradigm for technological innovation in literary translation. In J. L. Hadley, K. Taivalkoski-Shilov, C. Teixeira, and A. Toral (eds.), Using technologies for creative-text translation, pp. 18-39. New York: Routledge.

Rob Voigt, and Dan Jurafsky. 2012. Towards a Literary Machine Translation: The Role of Referential Cohesion. In D. Elson, A. Kazantseva, R. Mihalcea, and S. Szpakowicz (eds.), NAACL-HLT Workshop on Computational Linguistics for Literature, pp. 18-25. https://aclanthology.org/W12-25.pdf.

Stephen D. Richardson. 2012. Using the Microsoft Translator Hub at The Church of Jesus Christ of Latter-day Saints. In Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Commercial MT User Program. https://aclanthology.org/2012.amta-commercial.14.pdf.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In J. Hadley, M. Popović, H. Afli, and Andy Way (eds.), Proceedings of The Qualities of Literary Machine Translation, pp. 19-23. https://aclanthology.org/W19-7301.pdf.