Neural Machine Translation of Literary Texts from English to Slovene

Taja Kuzman
Department of Translation Studies
Faculty of Arts
University of Ljubljana
kuzman.taja@gmail.com

Špela Vintar
Department of Translation Studies
Faculty of Arts
University of Ljubljana
spela.vintar@ff.uni-lj.si

Mihael Arčan
Insight Centre for Data Analytics
Data Science Institute
NUI
Galway, Ireland
mihael.arcan@insight-centre.org

Abstract

Neural Machine Translation has shown promising performance in literary texts. Since literary machine translation has not yet been researched for the English-to-Slovene translation direction, this paper aims to fulfill this gap by presenting a comparison among bespoke NMT models, tailored to novels, and Google Neural Machine Translation. The translation models were evaluated by the BLEU and METEOR metrics, assessment of fluency and adequacy, and measurement of the post-editing effort. The findings show that all evaluated approaches resulted in an increase in translation productivity. The translation model tailored to a specific author outperformed the model trained on a more diverse literary corpus, based on all metrics except the scores for fluency. However, the translation model by Google still outperforms all bespoke models. The evaluation reveals a very low inter-rater agreement on fluency and adequacy, based on the kappa coefficient values, and significant discrepancies between post-editors. This suggests that these methods might not be reliable, which should be addressed in future studies.

1 Introduction

Recent years have seen the advent of Neural Machine Translation (NMT), which has shown promising performance in literary texts (Moorkens et al., 2018; Toral and Way, 2018). Most research on neural literary translation focused on the comparison of statistical and neural models, whereas this paper is one of the first to present a comparison exclusively among NMT models, specifically between models adapted to novels and the mixed-domain Google Neural Machine Translation (GNMT) system, exploring whether adaptation to literary text leads to better performance of NMT systems. This is also the first research paper that investigates literary machine translation (MT) from English to the highly inflected and under-resourced Slovene language. The models are evaluated both with automatic evaluation methodologies, more precisely the BLEU and the METEOR metrics, and human evaluation methods, i.e. an assessment of fluency and accuracy, a measurement of the temporal dimension of post-editing effort and error analysis. Since the neural models are evaluated by multiple evaluation methodologies, we are able to compare evaluation methods, and determine whether they are efficient.

Our hypotheses were that all models adapted to literary texts would yield better results than GNMT, based on automatic (hypothesis 1), as well as human evaluation (hypothesis 2), and that the model trained on out-of-domain parallel data and retrained on the novel *Practice Makes Perfect* (model ‘Novel’) would perform better than the model trained on out-of-domain parallel data and retrained on the corpus SPOOK (model ‘SPOOK’), according to both automatic (hypothesis 3) and human evaluation (hypothesis 4).

2 Related work

2.1 Machine translation of Slovene

The Slovene language poses challenges for MT due to its morphological complexity for all word classes and the lack of resources. Moreover, it is highly inflected, and it has a free word order (Krek, 2012). Nevertheless, several MT systems have been built between English and Slovene in recent times. In 2002, the first Slovene commercial MT system called Presis was developed (Romih and Holozan, 2002). This rule-based machine translation system was later followed by other foreign commercial systems, such as Bing Translator, Google Translate, Yandex Translate and Tradukka (Hari, 2018).

Additional systems were developed as a part of research projects, such as a statistical machine translation (SMT) system for Slovene subtitles, built in the framework of the SUMAT project (Etchegoyhen et al., 2014). Arčan et al. (2016) developed a publicly available mixed-domain SMT system called Asistent for translation between English and South Slavic languages, i.e. Slovene, Croatian and Serbian.

First comparisons of the performance of SMT and NMT approaches between English and Slovene were conducted in 2018, where SMT methods still outperformed NMT (Arčan, 2018). The translation quality of the NMT system can, however, be improved, by the addition of a parallel corpus containing selected sentences and by the enlargement of the neural architecture. Research, conducted by Donaj and Sepesy Maučec (2018), yielded more promising results. It revealed that NMT approach outperformed SMT in both English-to-Slovene and Slovene-to-English translation directions. Regarding the performance of commercial NMT systems for the translation between English and Slovene, Vintar (2018) compared Google's SMT and NMT for translating scientific texts with special focus on terminology translation. According to the BLEU score, GNMT outperformed the statistical system for both translation directions, however not for the translation of terms. In another study Hari (2018) compared the quality of Slovene translations of the English subtitles for the movie *The Lord of the Rings*, generated by the Bing Translator, GNMT and Yandex Translate. He discovered that Bing Translator outperformed GNMT and Yandex Translate.

2.2 State-of-the-Art in MT of Literary text

Until recently, there has not been much interest in the Computational Linguistics community regarding MT of literary texts, as the predominant opinion was that MT systems could never be useful for translating this type of text. Some of the first experiments were conducted in 2010 when Genzel et al. (2010) translated poetry with SMT systems from French to English and Greene et al. (2010) from Italian to English, producing translations that obey meter and rhyming rules. Another piece of research on literary machine translation from French to English was carried out by Jones and Irvine (2013), who translated samples of French prose and poetry using general-domain MT systems. Besacier (2014) conducted a post-editing experiment on SMT of literary texts from English to French which revealed that post-editing a pre-translated literary text could be used instead of a translation from scratch, although it does not achieve the same level of quality.

Toral and Way (2015) researched SMT of literary texts from Spanish to Catalan and carried out a human evaluation of the SMT models used. The findings revealed that evaluators considered 60% of the segments to be of comparable quality to professional human translation. In 2018, the same authors developed English-to-Catalan SMT and NMT models, tailored to literary texts, and compared them based on automatic and human evaluation. Both methods showed that the NMT system performed better, resulting in an 11% relative improvement over the SMT system (Toral and Way, 2018). Moorkens et al. (2018) also compared SMT and NMT systems, adapted for the translation of literature from English to Catalan, measuring post-editing effort with six participants. The findings revealed that all participants post-edited the NMT most quickly and that translation from scratch proved to be the most time-consuming. Moreover, the NMT model produced more fluent and adequate translations than the SMT one.

2.3 Analysis of Evaluation Methods

As manual evaluation is time-consuming and expensive to perform, it is regarded to be more accurate than automatic evaluation. However, research conducted by Callison-Burch et al. (2007) revealed low inter-annotation agreement for the assessment of fluency and adequacy, calling this method into question. To determine the inter-annotator agreement, they calculated the kappa coefficient, which is the proportion of time two or more annotators assigned identical scores to the same segments. According to Landis and Koch (1977), result from 0.0 to 0.2 means slight agreement, 0.21 to 0.4 fair, 0.41 to 0.6 moderate, 0.61 to 0.8 substantial and a higher score than 0.8 means almost perfect agreement. Analysis performed by Callison-Burch et al. (2007) revealed that the inter-annotation agreement for assessing fluency and adequacy was merely fair.

3 Experimental setup

In this section, we give an overview of the training and test datasets used in our experiment. Then, we present NMT systems and give insights into evaluation methods.

3.1 Training and Test Data

Bespoke models were trained on in-domain parallel data, either on the *Slovene Translation Corpus* (SPOOK) or on a corpus, consisting of a novel *Practice Makes Perfect*, written by Julie James, and its translation. In addition to these corpora, some models were also trained on out-of-domain parallel data to increase the lexical coverage of the training corpus. The out-of-domain data was mostly obtained from the OPUS web site (Tiedemann, 2012), which offers various parallel corpora, including Europarl, DGT, EMEA, KDE and EBC.

The *Slovene Translation Corpus* (SPOOK), a multilingual cross-comparable corpus of original and translated texts, was built in the framework of the Slovene Translation Studies: Resources and Research national research project which ran from 2009 to 2012. The corpus contains parallel corpora of literary texts in English, French, Italian and German and their translations to Slovene, as well as some original Slovene literary texts (Vintar, 2013). In this experiment, we used an English subcorpus consisting of nine English novels and their Slovene translations, i.e. J.R.R. Tolkien’s *Lord of the Rings: The Two Towers*, Dan Brown’s *The Da Vinci Code*, Eoin Colfer's *The Supernaturalist*, Colin Dexter's *The way through the woods*, Mark Haddon's *The Curious Incident of the Dog in the Night-Time*, Doris Lessing's *The Fifth Child*, J. K. Rowling's *Harry Potter and the Half-Blood Prince* and *Harry Potter and the Deathly Hallows*, and Zadie Smith's *White Teeth*. In total, it contains around one million English tokens.

In addition to that, we built a parallel corpus, consisting of Julie James’s romance novel *Practice Makes Perfect* and the Slovene translation *Osem let skomin*, produced by Irena Furlan. The corpus, built with the CAT tool MemoQ, consists of 7,000 segments and around 100,000 English tokens.

The test data was drawn from a similar corpus, consisting of a romance novel *Something about you* by Julie James, and its Slovene translation *Nekaj na tebi* by Irena Furlan. Thus, all models were tested on a novel by the same author and translated by the same translator as the novel on which our author-specific model *Novel* was trained. The dataset used for automatic evaluation consists of 2,547 segments and 41,054 English tokens. Since human evaluation is more time-consuming, participants in the experiment were given much shorter excerpts from the novel. Half of them were to post-edit and evaluate an excerpt *The Discovery of Body*, consisting of 16 sentences and 175 English words, and to translate from scratch an excerpt *The Interrogation*, containing 15 sentences and 174 words. For the other half the task was reversed: post-edit and evaluate the excerpt *The Interrogation* and translate the excerpt *The Discovery of Body*. For the purposes of error analysis, we analyzed MT outputs of these two excerpts and an excerpt from the beginning of the novel. The total length of the text that was analyzed is 929 words.

	Tokens		**Types**	
	English	**Slovene**	**English**	**Slovene**
Generic	62,067,541	5,1428,154	387,259	641,726
Spook	1,009,551	946,728	33,207	73,446
Practice	101,118	94,923	6,323	10,391
Something	41,054	39,014	3,895	6,215

Table 1. Statistics on datasets, used for training the neural translation models

3.2 MT systems

Google Neural Machine Translation is an NMT system, developed by Google in 2016. It supports 91 languages, including Slovene. Moreover, GNMT enables translation between language pairs never seen explicitly by the system, also known as “Zero-Shot Translation”. GNTM learns from millions of examples, which is made possible by Google’s machine learning toolkit TensorFlow and Tensor Processing Units (TPUs) (Schuster et al., 2016; Le and Schuster, 2016). Google’s current Universal Transformer NMT system is based on the standard Transformer, which is based on a self-attention mechanism and was found to outperform recurrent and convolutional models for English-to-German and English-to-French translation directions (Uszkoreit, 2017). In contrast to RNN-based approaches, the Universal Transformer processes all symbols at the same time and refines its interpretation by processing every symbol in parallel over multiple recurrent processing steps while making use of self-attention mechanism and devoting more attention to ambiguous words (Gouws and Dehghani, 2018).

Bespoke NMT models were trained using OpenNMT (Klein et al., 2017), a generic deep learning framework mainly specialized in sequence-to-sequence modelling. To improve the lexical coverage of out-of-vocabulary compound words, our NMT models were trained on sub-word units (Byte Pair Encoding). Initially, we used the default OpenNMT parameters, i.e. 2 layers, 500 hidden bidirectional Long Short-Term Memory (LSTM) units, 500 nodes, input feeding enabled, batch size of 64, 0.3 dropout probability and a dynamic learning rate decay. The networks were trained for 13 epochs. Then we also conducted some experiments by enlarging the neural architecture to 4 layers, 600 and 1,000 hidden LSTM units, and 600 and 1,000 nodes. As the results showed that the enlargement of the network did not have a large impact on the translation quality and that in some cases resulted in a decrease of the translation quality, we continued the training of the models with the default OpenNMT parameters. Similarly, experiments in which we trained the networks for up to 50 epochs did not result in the improvement of the translation quality, so we resumed the training of all models for 13 epochs.

In addition to GNMT and the generic NMT model (the baseline), trained on out-of-domain data, we evaluated multiple bespoke models, tailored to literature:

- model, trained on the corpus SPOOK (model 'Just SPOOK')
- model, trained on the novel *Practice Makes Perfect* (model 'Just Novel')
- model, trained on out-of-domain data and retrained on the corpus SPOOK (model 'SPOOK')
- model, trained on out-of-domain data and retrained on the novel *Practice Makes Perfect* (model 'Novel')
- model, trained on out-of-domain data and retrained on the corpus SPOOK and the novel *Practice Makes Perfect* (model 'SPOOK + Novel')

3.3 Evaluation

Firstly, all models were evaluated based on automatic evaluation methodologies. Then, we conducted a more detailed human evaluation of GNMT and two bespoke models, i.e. the SPOOK and the Novel NMT models. For the automatic evaluation, we used the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) metrics, which are based on the correspondence of the MT output and the reference translation. The BLEU score was obtained with the Interactive BLEU score evaluator,[1] which is available on the Tilde platform, whereas the METEOR score was calculated by the automatic machine translation evaluation system METEOR, available on GitHub.[2]

The human evaluation consisted of error analysis of the MT output, an assessment of fluency and adequacy, and a measurement of the temporal dimension of post-editing (PE) effort. Twelve Master's students in translation or interpreting took part in the evaluation. On average, participants had at least four years of translation experience and 83% of them have already had some PE experience. Each translated one excerpt from the novel *Something about you* by Julie James and post-edited the hypotheses of a similar excerpt, while assessing the fluency and adequacy of each segment. The translators were divided into six groups of two: groups A and B evaluated GNMT, C and D evaluated the translations provided by the SPOOK neural model, and E and F by the Novel model. In that way, all three models were evaluated by four participants each and on two excerpts. Participants also provided feedback after the translation via a questionnaire.

Participants translated and post-edited MT outputs using the Post-Editing Tool (PET) interface (Aziz et al., 2012), a CAT tool built for research purposes. PET measures time spent on editing each segment, tracks changes and allows adding optional assessments, which can be configured via a context file. Thus, after confirming a post-edited sentence, participants also assessed its fluency and adequacy on a pop-up assessment page before moving to the next sentence. Prior to the beginning of the assigned tasks, participants were provided with guidelines in order to produce professional quality translations. Moreover, they post-edited automatically generated translation of a short excerpt from the novel *Something about you*,

[1] https://www.letsmt.eu/Bleu.aspx

[2] https://github.com/cmu-mtlab/meteor

containing three sentences, to familiarize themselves with the PET tool and the workflow.

We followed TAUS guidelines for quality evaluation using adequacy and fluency approaches (Berghoefer, 2013). Participants were asked to rate adequacy on a 4-point scale based on the extent to which the meaning, expressed in the source, is also expressed in the MT output. Score 4 means that all meaning is expressed, 3 means most meaning, 2 little meaning and 1 means that no meaning is expressed in the hypothesis provided by the MT system. The second 4-point scale indicates how fluent and grammatically well-formed the hypothetical translation is. In this case, score 4 means that a translation is written in flawless Slovene, 3 means good Slovene, 2 means disfluent Slovene and 1 means that it is incomprehensible. After the assessment, we measured inter-annotation agreement using the kappa coefficient.

In addition to the measuring of the PE effort and assessing fluency and adequacy, we also compared GNMT, the SPOOK and the Novel NMT models based on an error analysis.

4 Results

4.1 Automatic Evaluation

Table 2 shows the results of the automatic evaluation. It revealed that GNMT achieved the best METEOR and BLEU score (30 and 21.97 respectively), followed by Novel with METEOR score of 20.35 and BLEU score of 20.75, and SPOOK with METEOR score of 19.67 and BLEU score of 19.01. These findings refute the first hypothesis predicting that models tailored to literature would achieve better scores than GNMT. On the other hand, the results confirmed the third hypothesis supposing that the Novel model, tailored to a specific author, would perform better than the SPOOK model, trained on a bigger but more varied literary corpus. The lowest score was obtained by the Just Novel model, with two layers. However, a similar model with four layers, trained on the same training set, obtained higher scores, although it produced considerably lower quality translations consisting of just six words. This indicates that BLEU and METEOR scores are not always accurate. The combined SPOOK + Novel model that was trained on the corpus SPOOK and on the corpus, consisting of a novel *Practice Makes Perfect* and its translation, performed worse than the models, trained on just one of those corpora. According to the BLEU metric, it performed even worse than the model, trained solely on out-of-domain data. This contradicts the common belief that the addition of more training data always leads to better results. In the case of the SPOOK + Novel neural model we can also observe a discrepancy between the BLEU and METEOR metrics. According to the METEOR metric, this model outperforms the baseline by 0.62 point, whereas based on the BLEU metric, it achieves 1.48 fewer points. Furthermore, the biggest difference between BLEU and METEOR scores is 8.03 points in the case of GNMT, whereas in the case of another model, the difference is only 0.40 point.

	Baseline	*Just SPOOK (2 layers)*	*Just SPOOK (4 layers)*	*Just Novel (2 layers)*	*Just Novel (4 layers)*	*SPOOK*	*Novel*	*SPOOK + Novel*	*GNMT*
BLEU	17.50	6.61	2.04	1.73	1.78	19.01	20.75	16.02	21.97
METEOR	18.50	11.86	6.98	5.01	5.21	19.67	20.35	19.12	30.00

Table 2. Results of the automatic evaluation

4.2 Measuring Post-Editing Effort

Since the time required for translation and post-editing varied among participants, the models were compared based on the time gains of post-editing. Nevertheless, the evaluation revealed significant discrepancies between post-editors. Table 3 illustrates that the first participant from the group C finished the translation task 7.4 minutes faster than the post-editing task, whereas the second participant from the same group finished the translation task 7.1 minutes slower than the post-editing task. This means that based on the second participant the evaluated model outperforms the other two, whereas based on the first participant, who post-edited the same output, the evaluated model performs the worst. Post-editors already had some experience in PE, they were given guidelines, and they had to post-edit a short excerpt before the evaluation. Therefore, the reason for the discrepancies between post-editors cannot be due to the lack of experience. It is probable that poor results can be attributable to the lack of precision and motivation. It is nonetheless true that no participant had more than 160 hours of PE experience–the equivalent of a month of full-time post-editing–which greatly increases the level of comfort with post-editing (Vasconcellos, 1986). In spite of discrepancies, the findings show that all three NMT approaches resulted in increases in translation productivity. In general, post-editing was revealed to be 1.6% faster than translation from scratch and most participants post-edited a pre-translated excerpt faster than they translated a similar excerpt. Based on the average times of all participants that assessed the same NMT model, the productivity increased the most in the case of GNMT, followed by the Novel and the SPOOK NMT models, as illustrated in Table 4. Most participants perceived post-editing to be faster than translation from scratch, although the perceptions of half of the participants did not match the measurements (highlighted bold in Table 3). Two out of three participants who finished the translation task faster than the PE task wrongly perceived the translation task to be more time-consuming.

Participants perceived the quality of outputs to be overall good or sometimes good. Their answers to the questionnaire revealed that most of them have positive attitudes towards post-editing. They mostly think that MT is more useful in assisting with professional translations of other types of text than literary texts, although some of them believe that might change in the future.

	Group A (person 1)– GNMT	*Group A (person 2)– GNTM*	*Group B (person 1)– GNMT*	*Group B (person 2)– GNMT*	*Group C (person 1)– SPOOK*	*Group C (person 2)– SPOOK*	*Group D (person 1)– SPOOK*	*Group D (person 2)– SPOOK*	*Group E (person 1)–Novel*	*Group E (person 2)–Novel*	*Group F (person 1)–Novel*	*Group F (person 2)– Novel*
Translation time (min)	9.5	12.6	6.0	12.3	17.2	8.4	12.8	14.0	11.7	17.6	10.7	12.1
PE time (min)	12.8	13.0	9.7	16.3	9.8	15.6	13.2	15.9	13.0	21.4	10.6	11.1
Difference between translation and PE time (min)	3.2	0.4	**3.8**	**4.0**	**-7.4**	7.1	**0.4**	1.9	1.3	**3.8**	-0.1	**-1.0**
The task that participants perceived to be more time-consuming	translation	translation	**PE**	**PE**	**translation**	translation	**PE**	translation	translation	**PE**	PE	**translation**

Table 3. Measurement of the temporal dimension of post-editing effort

	GNMT	*SPOOK*	*Novel*
Average difference between translation and PE time (min)	2.9	0.5	1.0

Table 4. Average difference between translation and PE time

4.3 Assessment of Fluency and Adequacy

Based on the assessment of fluency and adequacy, GNMT produced translations of the highest quality, followed by translations provided by the Novel neural model. However, the translations generated by the SPOOK model were given better scores for fluency. The results refute the second hypothesis predicting that models, tailored to literature, would achieve better scores than GNMT. On the other hand, the fourth hypothesis was partially confirmed, since the author-specific model performed better than the model, trained on a mixed literary corpus, according to the temporal dimension of post-editing effort and the assessment of adequacy. However, it obtained lower scores for fluency.

Figure 1 illustrates that not much can be inferred from the participants' assessments of fluency and adequacy. For instance, based on the assessment of the first participant from the group A, we could say that the GNMT produces the most fluent outputs. On the other hand, based on the assessment of the second participant from the same group we could infer that the GNMT's generated translations are the least fluent ones.

Inter-rater agreement on fluency and adequacy proved to be very low. Each hypothesis was evaluated by two participants. In two groups one sentence obtained the highest score in one or both categories by one evaluator and the lowest score by the other. In five out of six groups, one or more sentences were given the second-highest score by one evaluator and the lowest score by the other. In some cases, we can presume that the lowest score was given by mistake, since the evaluator decided that no post-editing is necessary for that segment. In other cases, the low-annotator agreement may be attributable to the issue that there are no clear guidelines on how to assign values to translations.

Inter-rater agreement was also measured using the kappa coefficient. The results revealed mostly slight inter-agreement. In group A, even a negative value occurred in one of the categories, whereas in the other category the inter-annotation agreement of the two participants was moderate, as shown in Figure 2 below.

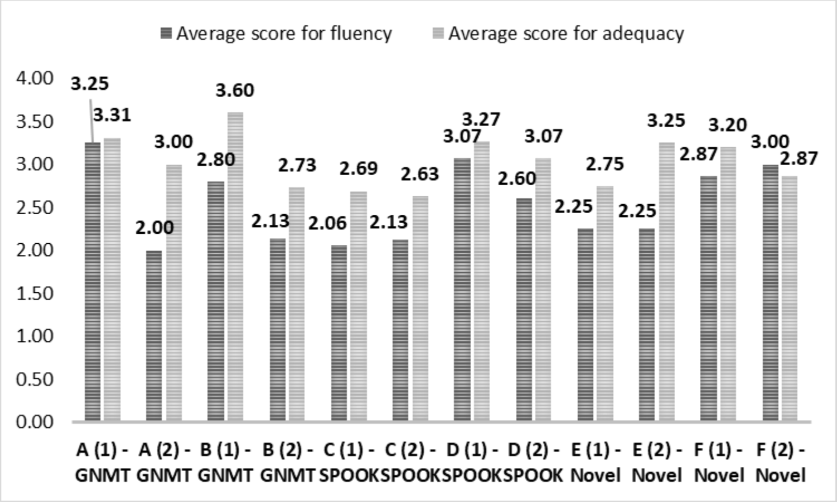

Figure 1. Average score for fluency and adequacy

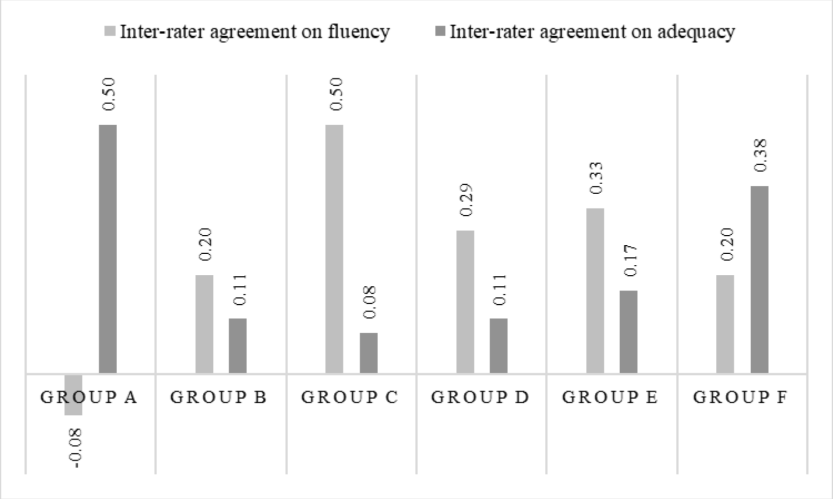

Figure 2. Inter-rater agreement on fluency and adequacy based on the kappa coefficient

4.4 Error Analysis

The error analysis of the translations generated by the GNMT, the Novel and the SPOOK models revealed various punctuation errors, wrong translations of prepositions and conjunctions, inappropriate shifts in verb mood, wrong noun forms and co-reference changes. Regarding semantic errors, the analysis revealed that GNMT assigned the wrong gender to the main character (‘Cameron’), the Novel model changed the name of another character, and all three models wrongly translated a proper noun of a hotel (‘Peninsula’) as a common noun. Many other semantic errors were detected, especially in connection with idioms and ambiguous words. Some expressions, such as “brunch buffet”, were inconsistently translated and the analysis revealed that when MT systems encounter a new word, GNMT most often leaves the term untranslated, whereas the SPOOK NMT model is especially prone to inventing words, which do not exist in the Slovene language. In addition to this, all models tend to omit and add words. The analysis revealed that the SPOOK and Novel neural models added or omitted negations, which significantly changes the meaning of the sentence. They also changed numbers, which can be perceived as a serious error in some cases. However, they also changed the American emergency number (911) to the Slovene emergency number (112), which can be perceived as a cultural adaptation. Nevertheless, such attempts can be problematic. For example, the Novel translation model substituted an imperial unit for the metric unit without converting the values, which led to an error.

The outputs of all three NMT models include some unintelligible sentences, as well as some sentences with only punctuation errors. However, there were no sentences that would not need post-editing.

5 Conclusion and Future Work

The automatic and human evaluation revealed that mixed-domain NMT model GNMT, trained on millions of examples, performs better than our models tailored to literature and trained on a much smaller training dataset. However, contrary to popular belief, more data does not always lead to better results, since the Novel NMT model, adapted to a specific author and trained on out-of-domain data and a corpus, consisting of one novel and its translation, outperformed the SPOOK one, trained on out-of-domain data and a bigger corpus, consisting of nine novels, written by various authors. Moreover, the model that was trained on the out-of-domain corpus and on both in-domain corpora performed worse than a model, trained solely on out-of-domain corpus, that is trained on a smaller training dataset. Since the Novel model, adapted to a specific author, came very close to the GNMT translation system based on the BLEU scores, future studies could fruitfully explore this issue further by training the model with more novels written by the same author. In our case, there are seven other novels by Julie James translated to Slovene that could be added to the training dataset.

In general, post-editing was revealed to be 1.6% faster than translation from scratch and most participants post-edited an excerpt faster than they translated a similar excerpt, which are promising results for literary machine translation from English to Slovene.

Moreover, the findings suggest that the assessment of fluency and adequacy and measurement of the temporal dimension of post-editing effort might not be reliable as evaluation methods. This assumption could be addressed in future studies which could be conducted on a larger scale, with more participants, preferably more experienced in post-editing, who would perform the task in a professional setting.

6 Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight), co-funded by the European Regional Development Fund.

References

Arčan, Mihael. 2018. A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. 3–10.

Arčan, Mihael, Maja Popović, and Paul Buitelaar. 2016. Asistent – A Machine Translation System for Slovene, Serbian and Croatian. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. 13–20.

Aziz, Wilker, Sheila Castilho M. de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. *The Eighth International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey. 3982–3987.

Berghoefer, Karin. 2013. *TAUS Best Practice Guidelines Quality Evaluation using Adequacy and/or Fluency Approaches*. URL: https://www.taus.net/index.php?option=com_rsfiles&layout=preview&tmpl=component&path=Articles%2Ftaus-adequacy-fluencyguidelines-may2013.pdf.

Besacier, Laurent. 2014. Traduction automatisée d'une oeuvre littéraire: une étude pilote. *Traitement Automatique du Langage Naturel (TALN)*, Marseille, France. 389–394.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation*. 136–158.

Denkowski, Michael, and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Donaj, Gregor, and Mirjam Sepesy Maučec. 2018. Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina. *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 62–68. Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, Iceland. 46–53.

Genzel, Dmitriy, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*. 158–166.

Gouws, Stephan, and Mostafa Dehghani. 2018. Moving Beyond Translation with the Universal Transformer. *Google AI Blog*, Google, 15 August 2018, https://ai.googleblog.com/2018/08/moving-beyond-translation-with.html.

Greene, Erica, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. *Proceedings of the 2010 conference on empirical methods in natural language processing.* 524–533.

Hari, Daniel. 2018. *Pregled prosto dostopnih strojnih prevajalnikov*. Thesis, University of Maribor.

Jones, Ruth, and Ann Irvine. 2013. The (un)faithful machine translator. *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* 96–101.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Opensource toolkit for neural machine translation.

Krek, Simon. 2012. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series. Springer.

Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. B*iometrics*, 33:159–174.

Le, Quoc V., and Mike Schuster. 2016. A Neural Network for Machine Translation, at Production Scale. *Google AI Blog*, Google, 27 September 2016, https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

Moorkens, Joss, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces* 7(2): 240–262.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*. 311–318.

Romih, Miro, and Peter Holozan. 2002. Slovensko-angleški prevajalni sistem (a Slovene-English translation system). *Proceedings of the 3rd Language Technologies Conference*, Ljubljana, Slovenia.

Schuster, Mike, Melvin Johnson, and Nikhil Thorat. 2016. Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. *Google AI Blog*, Google, 22 November 2016, https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

Tiedemann, Jörg. 2012. Character-based pivot translations for under-resourced languages and domains. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France. 141–151.

Toral, Antonio, and Andy Way. 2018. What Level of Quality Can Neural Machine Translation Attain on Literary Text?: From Principles to Practice. *Translation Quality Assessment*, 263–287. Springer, Cham, Switzerland.

Toral, Antonio, and Andy Way. 2015. Translating Literary Text between Related Languages using SMT. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, Colorado, USA. 123–132.

Uszkoreit, Jakob. 2017. Transformer: A Novel Neural Network Architecture for Language Understanding. *Google AI Blog*, Google 31 August 2017, https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html.

Vasconcellos, Muriel. 1986. Post-editing On-screen: Machine Translation from Spanish into English. *Proceedings of Translating and the Computer* 8, London, UK. 133–146.

Vintar, Špela. 2018. Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English-Slovene Language Pair. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Vintar, Špela. 2013. Uvodnik: o rojstvu korpusa SPOOK in njegovih prvih sadovih. *Slovenski prevodi skozi korpusno prizmo*, 6–13. Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia.