The Challenges of Using Neural Machine Translation for Literature

Evgeny Matusov
AppTek
Aachen, Germany
ematusov@apptek.com

Abstract

In this paper, we adapt state-of-the-art neural machine translation (NMT) systems to literary content and use them to translate fiction stories from English to Russian and from German to English. We show that such adapted systems have richer vocabulary and lead to improved automatic evaluation metrics on literary prose as compared to general domain NMT systems, including Google's online MT. We propose a new error classification scheme for NMT output that is specifically tailored to literary translation and let a bilingual evaluator analyze translated excerpts from two fiction stories. The results show that up to 30% of machine-translated sentences have acceptable quality. We observe very few severe syntactic errors even on complex sentences, but the meaning errors for ambiguous words are still numerous. A separate classification of consistency, pronoun resolution, and tone/register error types reveals a high potential of MT quality improvement by considering the context of previous sentences or even the whole story. A preliminary experiment aimed at reducing pronoun translation errors confirms this potential.

1 Introduction

Recent advances in neural machine translation led to a greater acceptance of MT technology, even among professional translators. However, it is hard to find anyone who would dare to use NMT for the professional translation of literature. Yet we believe that the challenges of literature translation could be tackled with NMT.

In this work, we adapted a baseline general domain NMT system, described in Section 3, to the style and diverse vocabulary of literary translations. The details of the adaptation process are given in Section 4. This was carried out on two language pairs: English-to-Russian and German-to-English. We then computed automatic error measures for translations of entire short novels and were able to show improvements as compared to the baseline model and Google's online MT (Section 5).

Next, we performed a thorough manual evaluation of both human and automatic translation quality on an excerpt from each novel. For better insights into the shortcomings of NMT and potential improvements, we devised a novel error classification scheme, as described in Section 6.2, intended to tackle errors characteristic of neural MT systems, including cohesion and inter-sentence context issues which are prominent in literary translation. Sections 6.3 and 6.4 describe these experiments in detail and also provide a quantitative comparison between Google's online and AppTek's adapted NMT for each error type.

We conclude the paper with a discussion on the possible applications of NMT for literature and underline the challenges, but also the opportunities associated with state-of-the-art NMT technology and its future developments.

2 Related Work

Using MT for literary translation has been inconceivable not only to professional translators of prose, but also to most MT researchers. As the technology made significant progress in the last decade, initial research in this direction appeared, although non-computational linguists remain largely skeptical (Almahasees and Mustafa, 2017). Voigt and Jurafsky (2012) identified that incorporating discourse features above the sentence level is an important requirement for literary translation because of the greater referential cohesion of literary texts, but did not run any MT experiments with systems adapted to such content. In a pilot study, Besacier and Schwartz (2015) trained a phrase-based statistical MT system for translating a short story from English to French, concluding that a faster literary translation with post-editing can be achieved at the expense of translator creativity and freedom of expression. Toral and Way (2018) compared phrase-based statistical MT with neural MT when translating literary content using automatic and human evaluation. They concluded that neural MT significantly outperforms phrase-based SMT in this genre, but "fills the gap" to the human quality level only by 20%. In that work, a vanilla NMT architecture for English-to-Catalan MT was used, not described in detail. The authors built a relatively large in-domain parallel corpus of human-translated fiction, and also use synthetic parallel data, for which Catalan novels are translated using a phrase-based system into English. In contrast, in our work we use the latest and best NMT architecture both for back-translation of large volumes of novels, and for the actual MT experiments; with only a very small parallel fiction corpus we are still able to obtain improvements over a strong general-domain NMT baseline.

Other related work important to literary translation include style transfer (Korotkova et al., 2018) and personalization (Rabinovich et al., 2016), number and gender disambiguation (Moryossef et al., 2019), document-level translation (Wang et al., 2017).

This work focuses on translation of prose; however, there have also been attempts to automatically translate poetry, with rhyming and rhythmical constraints, starting from the seminal work of Genzel et al. (2010) for phrase-based SMT. Recently, neural architectures were also proposed for this task (Ghazvininejad et al., 2018).

3 AppTek's Neural Machine Translation System

AppTek's NMT system is based on the the RETURNN toolkit (Zeyer et al., 2018) that implements training and inference in TensorFlow (Abadi et al., 2015). We trained two different architectures of NMT models: an attention-based RNN model similar to (Bahdanau et al., 2015) with additive attention for English-to-Russian and a Transformer model (Vaswani et al., 2017) with multi-head attention for German-to-English.

In the RNN-based attention model, both the source and the target words are projected into a 620-dimensional embedding space. The models are equipped with 4 layers of bidirectional encoder using LSTM cells with 1000 units. A unidirectional decoder with the same number of units was used in all cases. We applied a layer-wise pre-training scheme that lead to both better convergence and faster training speed during the initial pre-train epochs (Zeyer et al., 2018).

In the Transformer model, both the self-attentive encoder and the decoder consist of 6 stacked layers. Every layer is composed of two sub-layers: an 8-head self-attention layer followed by a rectified linear unit (ReLU). We applied layer normalization (Ba et al., 2016) before each sub-layer, whereas dropout (Srivastava et al., 2014) and residual connection (He et al., 2016) were applied afterwards. Our model is very similar to "base" Transformer of the original paper (Vaswani et al., 2017), such that all projection layers and the multi-head attention layers consist of 512 nodes followed by a feed-forward layer equipped with 2048 nodes.

We trained all models using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 for the attention RNN-based model and 0.0003 for the Transformer model. We applied a learning rate scheduling similar to the Newbob scheme based on the perplexity on the validation set for several consecutive evaluation checkpoints. We also employed label smoothing of 0.1 (Pereyra et al., 2017) for all trainings. The dropout rate ranged from 0.1 to 0.3.

AppTek's general domain English-to-Russian system was trained for roughly 3 epochs on 25 million sentence pairs (265M words on the English side). The corresponding German-to-English system was trained on 47M sentence pairs (752M running words on the English side) for less than 2 epochs.

4 Adaptation to Literary Content

First, AppTek's NMT had to be adapted to the style and diverse vocabulary of literary translations. In our experiments, we selected 2.3M sentences (23.5M running words) from books in Russian[1] and translated them using AppTek's general domain Russian-to-English NMT system. Following the approach of (Sennrich et al., 2016a), we then used the resulting parallel corpus as synthetic data, mixing it with the data that was used to train AppTek's general domain system from English to Russian. As parallel in-domain data, we used a small corpus of sentence-aligned texts[2] and the OPUS Books collection corpus[3] (Tiedemann, 2012), with a total of 270K sentence pairs and 5.2M running words on the English side.

We followed the same back-translation procedure for German-to-English, randomly selecting 10M sentences (155M running words) from English literature that we downloaded from the Gutenberg[4] project. Again, AppTek's highly competitive English-to-German general-domain Transformer model (Matusov et al., 2018) was used to translate these sentences, so that a synthetic parallel corpus could be used together with the other parallel data in NMT training of the reverse translation direction that was of interest to us. The in-domain parallel data consisted only of the small OPUS Books corpus with less than 50K sentence pairs and ca. 1.2M words on the English side.

We trained the system until convergence in terms of BLEU scores on held-out tuning data. For English-to-Russian, these were contiguous passages of Tolstoy's *Anna Karenina* (here, the input was the English translation, and Tolstoy's writing was used as the reference) and Chesterton's *The Innocence of Father Brown*. For English-to-German, the tuning set was the complete text of Kafka's *Der Prozess* from the OPUS Books collection.

5 Experimental Results

In this section, we review the automatic scores for the generated literature translations. We compute case-insensitive BLEU and TER scores (Papineni et al., 2002; Snover et al., 2006).

System	BLEU [%]	TER [%]
Google	13.9	84.6
AppTek	14.2	83.7
+ adaptation	15.2	82.5

Table 1: Automatic MT quality measurements for English-to-Russian literary translation.

System	BLEU [%]	TER [%]
Google	20.2	67.2
AppTek	18.5	69.7
+ adaptation	16.2	71.0

Table 2: Automatic MT quality measurements for German-to-English literary translation.

5.1 English-to-Russian

We evaluated the quality of Google's online MT, AppTek's general domain and literature-adapted NMT on four sentence-aligned stories by Conan-Doyle (*The Lift*, *Scandal in the Bohemia*) Poe, (*The Pit and the Pendelum*), and Chesterton (*The Invisible Man*). Thus, the test set was comprised of 1646 sentences and 30K words on the English side.

The experimental results are summarized in Table 1. First, we see that the BLEU scores are much lower than those of state-of-the-art systems on newswire and news commentary texts as evaluated e.g. at WMT 2019[5] (the BLEU scores there are mostly over 30%). This supports the assumption about the particular difficulty of literary translation, but, as we will discuss in Section 6.1, also highlights serious errors in human reference translation.

Google's online NMT and AppTek's baseline system both perform at similar level, with AppTek's system showing marginally better scores on literary content. AppTek's En-Ru system adapted to literary content improves over the general domain baseline by 1% BLEU absolute and thus also outperforms Google's online MT (15.2 vs. 13.9% BLEU). However, as we will see in Section 6.3, these score improvements do not necessarily mean better translation quality according to human analysis.

5.2 German-to-English

For German-to-English, the test set we chose was Franz Kafka's *Verwandlung* with 675 sentences and ca. 20K German words; interestingly, this cor-

[1] The books are publicly available from lib.ru and other sources.

[2] Crawled from http://multitran.ru.

[3] http://opus.nlpl.eu/Books.php

[4] https://www.gutenberg.org

[5] http://matrix.statmt.org/matrix/systems_list/1914

pus was also selected by (Cap et al., 2015) for their experiments on co-reference resolution in literary texts, where they argue that a co-reference resolution algorithm can be improved by features derived from word alignment to a human translation of the text into another language.

Table 2 summarizes the automatic error measures for Google's online MT, Apptek's general domain and adapted NMT. Google's system outperforms Apptek's systems for this language pair, but as the human analysis will show in Section 6.4, there were error categories, for which AppTek's output had less errors. The adaptation using back-translated English literature did not result in BLEU and TER score improvements, but again a bilingual evaluator confirmed that the output of the adapted system was better across multiple error categories. This underlines again that automatic MT error measures are not reliable for judging the quality of literary translation.

6 Error Analysis

We employed a bilingual evaluator fluent in the source and target languages to perform an error analysis of the MT output on parts of the test set, i.e. on the excerpts of Conan-Doyle and Kafka stories for English-to-Russian and German-to-English, respectively.

6.1 Human Translation Quality

Before dealing with MT output, the human expert thoroughly checked the human reference translations by comparing them to the source sentences.

To our surprise, the Russian human translation had a significant number of errors. Some of them (5 in total) could be explained by wrong automatic sentence alignment, where a part of the reference translation for a given segment actually was a translation of (a part of) the previous or the next segment. However, we also noticed other unexpected errors, including simplifications, omissions, and meaning change, which, in our opinion, go beyond the usual freedom of a translator to deviate from literal translation of the original text. Here are some examples:

- *Don't worry, my darling, the cloud will roll off.* is translated into Не волнуйся, дорогая, всё пройдёт [*don't worry, dear, all will pass*] which means that the translator could not find a good idiomatic equivalent and translated the idiom as "everything will pass".

- *Then it lifts quite suddenly, like a mist in the sunshine.* For this sentence, the translator completely reversed the meaning of the verb "lifts", translating it into появляется, "appears".

- The word *nightmare* is translated into ночной кошмар [*night nightmare*], an error that a professional translator can't afford to make.

- The term *side show* is omitted from the translation, perhaps because it was hard for the translator in Russia in the pre-Internet era to check what it means.

- *It's hung up, but the gear is being overhauled.* The sentence was translated as Немного задерживаемся, механизм осматривают. [*(We are) somewhat delayed, mechanism is being looked at.*] Here, the first part of the sentence is translated as "we are a bit delayed", although it is clear from the context that the gear is stuck, which has more consequences than a simple delay.

- *... a man who was descending the steel framework.* The phrase was translated as ... человек, который опускал вниз стальной каркас. [*... man, who brought down the steel carcass.*] Here, the translator thought that the man brought down the steel framework, although it is clear from previous and subsequent sentences that the man was climbing down the framework of the lift shaft.

Overall, there were 29 errors in 111 segments which significantly altered the meaning intended by the author and/or omitted translations of some words or phrases.

For Kafka's translation into English, the situation is somewhat better: here, we found only 7 errors, and only two segmentation errors. An example of a severe error is a translation of the sentence *Gregor war während seines fünfjährigen Dienstes noch nicht einmal krank gewesen*, which was translated into "in fifteen years of service Gregor had never once yet been ill", whereas actually Gregor was only employed for 5 years. Another error where the meaning is completely reversed was noted in the translation of the following segment: *Gregor erschrak, als er seine antwortende Stimme hörte, die wohl unverkennbar seine frühere war..* This was translated into "Gregor was shocked when he heard his own voice answering, it could hardly be recognised as the voice he had had before...".

6.2 MT Error Classification

In previous work, a number or MT error classification schemes have been proposed (Flanagan, 1994; Popović and Ney, 2011; Costa et al., 2015). All of them were either linguistically motivated or designed with the goal of identifying and classifying errors (semi-)automatically. After analyzing literature translation output, we have come to a different classification that specifically addresses higher-quality neural machine translation and highlights errors which can be fixed with additional context or information. We also introduce an idiom translation error category which is very important for literature.

Here are the proposed categories in detail:

1. *M1: severe meaning error.* A word or a short phrase is translated into a word or phrase in the target language with a wrong meaning given the context, and this translation is misleading to the reader. The reader can not easily recover the original meaning without seeing the source sentence. For NMT systems, in most cases these are ambiguous words or phrases, since wrong translations into something completely unrelated are rare, except for unknown/rare words, for which we introduce a separate category below.

2. *M2: minor meaning error.* A translated word or a short phrase conveys the original meaning that was intended in the source language, but with slight deviations. Usually, a synonym is used that has a slightly different meaning or is stylistically or otherwise not appropriate given the context. Yet the intent of the author can be understood from the translation and a better formulation can be guessed by the reader without consulting the source sentence.

3. *U: unknown word or segmentation error.* The vast majority of NMT systems use subwords (Sennrich et al., 2016b; Kudo and Richardson, 2018) to represent translation units. Thus, any out-of-vocabulary (OOV) word is separated into several known subwords. This does not guarantee a correct translation of the OOV word in any way. Moreover, known rare words may be translated incorrectly if the subwords of such a word have a pronounced different meaning. We group all such errors in a single category; in most cases these errors are directly visible in the MT output (e.g. wrong transliteration, translation of only a part of a word, etc.). For source languages without explicit word segmentation, such as Chinese, this category would also contain MT errors resulting from incorrect word segmentation.

4. *C: Consistency/term translation error.* This category specifically addresses translation consistency for words and phrases that, in the context of a particular document, should have a unique translation (apart from morphological variation) throughout the document. Examples include names and name transliterations, as well as technical or other terms (cf. *Flying Service* in Conan-Doyle's text and *Prokurist* in Kafka's text).

5. *P: pronoun resolution error.* As the MT quality improved with neural systems, these errors, which in many cases can be avoided only by consulting the context of the previous sentence(s) or even the whole document have become more visible, hence we introduced a separate category for them.

6. *L: locution error.* Whereas such errors could be categorized as meaning errors, we introduce a separate category for wrong locution or idiom translation. An idiom translation is considered wrong if the idiom is translated word-for-word, which significantly distorts its meaning in the target language, or into an idiom that has a different meaning or a similar meaning, but is incomplete/erroneously formulated.

7. *O, I, R: omission, insertion, repetition errors.* These three error categories have been frequently used in the MT community. Whereas, as our analysis shows, insertion errors (insertion of an unrelated word or phrase) are very rare in NMT output, omissions, i.e. untranslated word sequences, still happen, especially in longer sentences. Repetition errors include not only repetitions of single words or phrases, but repetitions with conjunctions (e.g. "'red and red") or repetitions in a differ-

ent word form “wooden wood” or constructs such as “doorbell door”.

8. *S1: severe syntax error.* The structure of the translated sentence is not correct. It can’t be parsed by a human, or the incorrect syntax distorts the meaning of the entire sentence, even though the meaning of individual words and short phrases is conveyed correctly. Examples include passive constructions with subject/object wrongly swapped, wrong tense, wrong attachment of prepositional phrases, morphological disagreement leading to parsing ambiguity, etc. To some extent, there is overlap with M1, but the S1 errors can not be easily localized to a single word/phrase.

9. *S2: minor syntax error.* The translated sentence contains minor syntactic or morphological errors, which can be easily corrected without significant changes to the sentence. Examples may include wrong verb tense without meaning distortion (e.g. simple vs. progressive), morphological agreement between noun and adjective, a not very appropriate preposition where a better one can be easily guessed, etc.

10. *T: tone/register error.* These errors may affect multiple words in a sentence, but only one error per sentence is counted. Examples include a wrong “you”-form and corresponding verb forms (polite vs. informal), word forms addressing a male when from previous context it is clear that a female should be addressed, etc. Another example is a formally correct translation of German “man kann” into English as “one can” or “you can”, which is in practice often not appropriate. Also, the usage of stylistically inappropriate words and phrases (e.g. colloquialisms) falls under this category.

6.3 English-to-Russian MT

Table 3 summarizes the results of the error analysis performed by a bilingual evaluator according to the error classification described in Section 6.2. The error analysis was performed separately for each of the systems analyzed (Google’s online MT and AppTek’s NMT adapted to literary content) on the first 114 segments of A. Conan-Doyle’s *The Lift*, which was part of the test set mentioned in Section 5.1. In Table 3, we also show the BLEU and TER scores on these segments only. The human expert had access to all 114 segments at once when marking/counting errors in the MT output. The 114 segments contained 1489 English words.

We observed that although BLEU and TER improvements of the AppTek’s adapted system are substantial, they are not reflected in human analysis. Approximately 20% of segments for both MT systems did not contain any errors (OK) and thus would not require any further processing by a professional translator or post-editor. AppTek’s MT output has fewer severe meaning errors (30 vs. 33) and fewer minor syntax errors (22 vs. 29). However, this comes at the expense of an increased number of minor meaning errors, where a wrong synonym is used (30 vs. 20). One can argue, however, that these errors by definition can be fixed by a monolingual post-editor of the target language.

Given the small sample size of 114 sentences, the number of consistency (C) and pronoun resolution (P), and tone/register errors (T) is rather high and suggests that document-level context is necessary to improve performance. For consistency errors, terminology override could be used to enforce e.g. that “the lift” is translated always as подъемник and not лифт, but it is an open research problem how to achieve this in morphologically rich target languages, where multiple word forms of the desired term translation may have to be produced (in this example, up to six different noun cases of подъемник).

The number of omission errors (O) is high (7 and 11), which supports previous findings about NMT errors. On the other hand, the number of serious syntax errors is low, which again supports the argument that NMT systems generally produce fluent and syntactically correct output. This also suggests that the post-editing required to fix the remaining errors would probably be local in most cases, where only single words or groups of words would have to be corrected, as opposed to re-structuring the entire sentence. A good example for such minimal post-editing is the following MT output for a complex sentence from one of the systems: Барнс, рабочий, пробормотал, что что-то должно быть не так, и прыгнул, как кошка, через щель, отделявшую их от решетки из металла, он вылез из поля зрения. The English sentence was: *Barnes, the workman, mut-*

System	BLEU	TER	OK	M1	M2	U	C	P	L	O	R	I	S1	S2	T	Total
Google	11.1	86.4	22	33	20	2	13	6	9	7	3	0	2	29	5	129
AppTek	13.6	80.8	23	30	30	2	12	10	11	11	4	0	4	22	10	146

Table 3: Human error analysis and BLEU and TER scores in % on the first 114 segments of A. Conan-Doyle's *The Lift* of Google's online MT and AppTek's literature-adapted NMT. The acronyms of the error categories are explained in Section 6.2.

tered that something must be amiss, and springing like a cat across the gap which separated them from the trellis-work of metal he clambered out of sight. Here, it is enough to fix one letter, changing the past tense verb прыгнул [*jumped*] into a gerund прыгнув [*jumping, springing*].

Finally, the high number of idiom translation errors (L) indicates a high number of idioms in the text by Conan-Doyle (mostly spoken by the characters of *The Lift*), and the inability of NMT systems to translate them. Here, idiom dictionaries could be of help, but unfortunately, they are rarely available in electronic form and are in most cases not used by MT system developers because of copyright issues.

6.4 German-to-English MT

Table 4 summarizes the results of the analysis by the bilingual human expert of the MT output for the first 114 segments of F. Kafka's *Die Verwandlung*. The 114 segments contained 2478 German words, which means that the sentence length here is on average 66% longer than for the English-to-Russian segments analyzed in the previous section. Nevertheless, the total number of errors is slightly lower for En-De than for En-Ru, which shows a higher level of MT quality for this language pair. Overall, 28-30% of the segments were considered as acceptable by the bilingual evaluator, which is also higher than for English-to-Russian.

Again, although the BLEU scores on these segments show that the Google online system is significantly better, the error analysis reflects this only in part. For instance, AppTek's output has no repetition or insertion errors, and fewer omission and severe syntax errors than Google's output. On the other hand, Google is somewhat better at meaning preservation (M1 and M2 errors).

The high quality of translations from German to English can be illustrated with multiple examples using sentences with complex structure, see Table 5. From the examples of AppTek's literature-adapted NMT, we can see that a richer vocabulary is used (e.g. the words "recollected", "alas", "enveloped", "clumsy"). In fact, we measured a 4% larger vocabulary in the AppTek's translation of *Die Verwandlung* as compared to Google's output.

To test whether pronoun resolution errors can be avoided by introducing the context of the previous source sentence, we trained a variant of the adapted model in which we joined two subsequent short German sentences from the same document with a special separator symbol, whenever the second sentence contained a pronoun and the total number of words in the joined sentence did not exceed 50. The joining was done also for the corresponding English sentences to make valid training sentence pairs. Such data was then added to the original training data. At translation time, we did the joining on the source side only, and then evaluated only the part of the MT output after the generated separator symbol. The result did not change the BLEU score significantly (it increased from 18.2 to 18.7), but two pronoun errors were corrected[6] for the following Kafka's text: *Sollte der Wecker nicht geläutet haben? [Should not the alarm-clock have been ringing?] Man sah vom Bett aus, daß er auf vier Uhr richtig eingestellt war; gewiß hatte er auch geläutet.* Here, if the second sentence is translated separately by the AppTek's literature-adapted system, the translation is "It was seen from the bed that he was properly set at four o'clock; certainly he had also ringed." Google's translation also makes similar pronoun translation errors: "From the bed you could see that he was right at four o'clock; he had certainly rung, too.". In contrast, our system that was additionally trained on joined pairs of sentences and also encoded the previous sentence, produced a much better output: "It was seen from the bed that it was set to four o'clock; surely it was ringing."

The preliminary experiment above showed that it is possible to benefit from inter-sentence context for literature translation. It remains to be seen what NMT architecture and, more importantly, evaluation criteria are most suitable for this endeavour.

[6] The training of the system in question finished too late for a full human analysis, so here we only looked at the sentences with previously identified pronoun resolution errors.

System	BLEU	TER	OK	M1	M2	U	C	P	L	O	R	I	S1	S2	T	Total
Google	22.9	64.0	36	25	22	2	11	9	4	17	3	1	6	23	2	125
AppTek	18.2	67.7	32	31	24	6	9	9	3	14	0	0	4	30	2	132

Table 4: Human error analysis and BLEU and TER scores in % on the first 114 segments of F. Kafka's *Die Verwandlung* of Google's online MT and AppTek's literature-adapted NMT. The acronyms of the error categories are explained in Section 6.2.

German source (F. Kafka)	Human Reference	Google's online MT	AppTek's adapted NMT
Er lag auf seinem panzerartig harten Rücken und sah, wenn er den Kopf ein wenig hob, seinen gewölbten, braunen, von bogenförmigen Versteifungen geteilten Bauch, auf dessen Höhe sich die Bettdecke, zum gänzlichen Niedergleiten bereit, kaum noch erhalten konnte.	He lay on his armour-like back, and if he lifted his head a little he could see his brown belly, slightly domed and divided by arches into stiff sections. The bedding was hardly able to cover it and seemed ready to slide off any moment.	He lay on his panzerartig hard back and saw, if he raised his head a little, his arched, brown, divided by arc-shaped stiffened stomach on the height of the blanket, ready for total descent, could barely maintain.	He lay on his armor-like hard back, and saw, when he lifted his head a little, his vaulted, brown belly, divided by bow-shaped stiffenings, on the height of which the duvet, ready for complete slipping, could scarcely yet be preserved.
In solchen Augenblicken richtete er die Augen möglichst scharf auf das Fenster, aber leider war aus dem Anblick des Morgennebels, der sogar die andere Seite der engen Straße verhüllte, wenig Zuversicht und Munterkeit zu holen.	At times like this he would direct his eyes to the window and look out as clearly as he could, but unfortunately, even the other side of the narrow street was enveloped in morning fog and the view had little confidence or cheer to offer him.	At such moments he aimed his eyes as sharply as possible at the window, but unfortunately, the sight of the morning mist, which even covered the other side of the narrow street, did not bring much confidence and cheerfulness.	At such moments he directed his eyes as sharply _ to the window, but, alas, from the sight of the morning mist, which even enveloped the other side of the narrow street, was to fetch little confidence and murmur.
Er erinnerte sich, schon öfters im Bett irgendeinen vielleicht durch ungeschicktes Liegen erzeugten, leichten Schmerz empfunden zu haben, der sich dann beim Aufstehen als reine Einbildung herausstellte, und er war gespannt, wie sich seine heutigen Vorstellungen allmählich auflösen würden.	He remembered that he had often felt a slight pain in bed, perhaps caused by lying awkwardly, but that had always turned out to be pure imagination and he wondered how his imaginings would slowly resolve themselves today.	He remembered having often felt in bed some slight pain, perhaps awkward, that turned out to be pure imagination when he got up, and he wondered how his present ideas would gradually dissolve.	He recollected having often felt some slight pain caused by clumsy lying in the bed, which then turned out to be pure imagination when getting up, and he was eager to see how his present notions would gradually dissolve.

Table 5: Examples of German-to-English NMT quality. Substantial MT errors are highlighted in red, good word and phrase choices in green.

7 Conclusions and Discussion

In this work, we challenged the assumption that MT is not suitable for literary translation. We adapted state-of-the-art neural MT systems for English-to-Russian and German-to-English to Russian and English fiction, respectively, by using back-translated data and observed that such adaptation leads to improved translation quality according to automatic evaluation metrics. We then asked a bilingual evaluator to thoroughly analyze the adapted MT output according to a novel error taxonomy tailored specifically to NMT errors and potential areas for improvement, with the following observations:

- Up to 30% of evaluated segments, mostly short sentences, were considered acceptable and might only require proof-reading by a monolingual editor of the target language.

- NMT of German fiction into English subjectively has higher quality than NMT of English literature into Russian; in fact the quality is often high enough to understand and even enjoy the story.

- Longer sentences are translated well in terms of syntactic structure, so that the necessary post-editing is often local and minor.

- Automatic evaluation using a single, often badly sentence-aligned human reference is unreliable; moreover, the human translation may contain severe meaning and other (e.g. omission) errors.

- There is significant potential to improve MT quality beyond genre adaptation by using inter-sentence context. This is especially true for consistent translation of character names, places, as well as pronoun resolution and translation style (e.g. formal vs. non-formal).

To conclude, we would like to elaborate on potential use cases of NMT for literature. Automatic translation of literature may be useful not only for helping professional literature translators in a post-editing scenario. It can also help to make largely undiscovered foreign language books instantly available online to readers worldwide, e.g. when they are translated into English. Publishers could also use NMT to better familiarize themselves with such foreign literary works and be aided in their selection process of books to professionally translate into another language, thus promoting an increased circulation of high-quality work among different languages and cultures.

Automatic translation of prose in combination with MT quality estimation methods could also be used to identify segments which are difficult to translate, or where there is a higher likelihood for a translator to make an error. Literary translations are rarely proof-read by bilinguals, but rather a monolingual editor of the target language edits the translation before publication, a process during which there is a risk of errors being introduced in the text. We argue that a higher level of quality control of literary translation is necessary, and NMT systems could prove to be useful tools to facilitate and speed up this process.

In another application, a good book translated by NMT with consistent name translations could facilitate its crowd-sourced translation (similarly to crowd-sourced subtitling for popular films and series), which could lead to improved quality of such fan translations. Finally, automatic translation may assist foreign language learners with specific phrases they have trouble understanding when reading a book in said foreign language, and thus NMT could have useful applications in foreign language learning as well.

Acknowledgements

The author would like to thank Yota Georgakopoulou for helpful feedback on the first version of this paper.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Almahasees, Z and Zakaryia Mustafa. 2017. Machine translation quality of Khalil Gibran's the Prophet. *AWEJ for translation & Literary Studies Volume*, 1.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*. Version 1.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May.

Besacier, Laurent and Lane Schwartz. 2015. Automated translation of a literary work: a pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122.

Cap, Fabienne, Ina Rösiger, and Jonas Kuhn. 2015. A pilot experiment on exploiting translations for literary studies on Kafka's "Verwandlung". In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 48–57, Denver, Colorado, USA, June. Association for Computational Linguistics.

Costa, Ângela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Flanagan, Mary. 1994. Error classification for MT evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 65–72.

Genzel, Dmitriy, Jakob Uszkoreit, and Franz Och. 2010. "poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA, October. Association for Computational Linguistics.

Ghazvininejad, Marjan, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May.

Korotkova, Elizaveta, Maksym Del, and Mark Fishel. 2018. Monolingual and cross-lingual zero-shot style transfer. *arXiv preprint arXiv:1808.00179*.

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Matusov, Evgeny, Patrick Wilken, Parnia Bahar, Julian Schamper, Pavel Golik, Albert Zeyer, Joan Albert Silvestre-Cerda, Adria Martınez-Villaronga, Hendrik Pesch, and Jan-Thorsten Peter. 2018. Neural speech translation at AppTek. In *International Workshop on Spoken Language Translation*.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. *arXiv preprint arXiv:1903.03467*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Pereyra, Gabriel, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548.

Popović, Maja and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Rabinovich, Ella, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. pages 86–96, August.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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*, pages 223–231, Cambridge, Massachusetts, USA, August.

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Voigt, Rob and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25.

Wang, Longyue, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *arXiv preprint arXiv:1704.04347*.

Zeyer, Albert, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 128–133.