Noname manuscript No.
(will be inserted by the editor)

A User-Study on Online Adaptation of Neural Machine Translation to Human Post-Edits

Sariya Karimova · Patrick Simianer · Stefan Riezler

Received: date / Accepted: date

Abstract The advantages of neural machine translation (NMT) have been extensively validated for offline translation of several language pairs for different domains of spoken and written language. However, research on interactive learning of NMT by adaptation to human post-edits has so far been confined to simulation experiments. We present the first user study on online adaptation of NMT to user post-edits. Our study involves 29 human subjects (translation students) whose post-editing effort and translation quality were measured on about 4,500 interactions of a human post-editor and a machine translation system integrating an online adaptive learning algorithm. Our experimental results show a significant reduction of human post-editing effort due to online adaptation in NMT according to several evaluation metrics, including hTER, hBLEU, and KSMR. Furthermore, we found significant improvements in BLEU/TER between NMT outputs and human references (from professional translators), and a strong correlation of these improvements with quality improvements of post-edits.

Keywords online adaptation · post-editing · neural machine translation

1 Introduction

The attention-based encoder-decoder framework for neural machine translation (NMT) (Bahdanau et al, 2015) has been shown to be advantageous over

Sariya Karimova · Patrick Simianer · Stefan Riezler
Heidelberg University - Department of Computational Linguistics
69120 Heidelberg, Germany
Tel.: +49622154-3245
E-mail: {karimova, simianer, riezler}@cl.uni-heidelberg.de

Sariya Karimova
Kazan Federal University
420008 Kazan, Russia

the well-established paradigm of phrase-based machine translation immediately after its inception. For example, significant improvements according to automatic and manual evaluation metrics could been shown in benchmark translation competitions for spoken language (Luong and Manning, 2015) and written language (Jean et al, 2015; Sennrich et al, 2016a). These results have been investigated in-depth in analyses of the advantages of NMT along linguistic dimensions (Bentivogli et al, 2016b) and along different domains (Castilho et al, 2017a). In contrast, research on uses of NMT in online adaptation scenarios has so far been confined to simulations where the interactions of an NMT system with a human post-editor are simulated by a given set of static references (Wuebker et al, 2016; Knowles and Koehn, 2016; Turchi et al, 2017; Peris et al, 2017, *inter alia*) or by a set of offline generated post-edits (Turchi et al, 2017). User-studies on the benefits of machine learning for adaptation of translation systems to human post-edits are rare, and to the best of our knowledge, such studies have so far been restricted to phrase-based machine translation (Denkowski et al, 2014b; Green et al, 2014; Bentivogli et al, 2016a; Simianer et al, 2016).

We present a user study that analyzes 4,500 per-sentence interactions between an NMT system and 29 human post-editors. Our target domain are patents that have to be translated from English into German. Our goal is to quantify the mutual benefits of a system that immediately learns from user corrections, on the one hand by reducing human post-editing effort, and on the other hand by improving translation quality of the NMT output. In comparing post-editing of NMT outputs that are generated from systems with and without online adaptation, we find a significant reduction in post-editing effort for the former scenario according to the metrics of hTER (and hBLEU) between NMT outputs and human post-edits. This confirms findings that have been reported for user studies on online adaptation of phrase-based systems (Bentivogli et al, 2016a; Simianer et al, 2016). Moreover, we find significant improvements of post-editing effort for the online adaptation scenario with respect to metrics such as keyboard strokes and mouse clicks that have been used in computer-assisted translation.

We also attempt to quantify improvements in translation quality by measuring improvements in sentence-level BLEU+1 (Nakov et al, 2012) and TER between the iteratively refined NMT outputs and static human reference translations. We find significant improvements with respect to both metrics, showing a domain adaptation effect due to online adaptation. This effect is propagated further in an improvement of BLEU and TER between post-edits created by translation students and reference translations produced by professional patent translators. Measuring BLEU/TER improvement of post-edits to references is a useful tool to measure student translators' improvements in domain-specific technical vocabulary and patent-specific constructions.

The remainder of this paper is organized as follows: In Section 2, we discuss the related work. We briefly introduce the learning protocol of online adaptation in Section 3. Then we describe the tools and data, and the experimental design of our user study (Section 4). Experimental results will be discussed

in Section 5. We conclude the paper by conclusions to be drawn from our experiments (Section 6).

2 Related Work

The advantages of NMT and its challenges have been investigated from different angles in recent work (Koehn and Knowles, 2017; Toral and Sánchez-Cartagena, 2017; Farajian et al, 2017; Macketanz et al, 2017; Castilho et al, 2017a; Klubička et al, 2017; Bentivogli et al, 2018; Isabelle et al, 2017; Popović, 2017; Forcada, 2017; Castilho et al, 2017b; Junczys-Dowmunt et al, 2016; Klubička et al, 2018; Shterionov et al, 2017; Burchardt et al, 2017; Menacer et al, 2017, *inter alia*), however, studies on interactive NMT, especially user studies involving human post-edits of NMT outputs, have so far not been presented.

Online adaptation has been thoroughly researched since at least a decade, either by adding online discriminative learning techniques (Cesa-Bianchi et al, 2008; Martínez-Gómez et al, 2012; López-Salcedo et al, 2012; Denkowski et al, 2014a; Bertoldi et al, 2014, *inter alia*) to phrase-based MT systems, or adaptations to the generative components of the phrase-based framework (Nepveu et al, 2004; Ortiz-Martínez et al, 2010; Hardt and Elming, 2010, *inter alia*). Recent studies applied the online adaptation framework to NMT, however, by simulating the interactive scenario by online learning from offline created human references or post-edits (Wuebker et al, 2016; Knowles and Koehn, 2016; Turchi et al, 2017; Peris et al, 2017).

User-studies involving the generation of post-edits in an online interaction between translation system and post-editor have been presented as well, however, to the best of our knowledge these studies have been confined so far to phrase-based MT (Green et al, 2013, 2014; Denkowski et al, 2014b; Bentivogli et al, 2016a; Simianer et al, 2016).[1] The closest approach to our work is the study presented by Bentivogli et al (2016a). Similar to our work, they involve human subjects in an interactive post-editing scenario where the system learns online from user corrections. However, their study is confined to phrase-based machine translation, and differs from our study in choosing a within-subjects experimental design where the same translator post-edits the same document under either test condition (static/adaptive NMT). We use a more standard between-subjects design where each session is composed of different documents of comparable difficulty, each of which is translated under two different conditions (with and without online NMT adaptation), and post-edited by two different translators.

[1] In addition, Green et al (2014) – one of the first user studies on online adaptation to post-edits – performed system updates offline instead of online.

Train global model M_g
for each document d of $|d|$ segments
  for each example $t = 1, \ldots, |d|$
    1. Receive input sentence x_t
    2. Output translation $\hat{y}_t$
    3. Receive user post-edit y_t
    4. Refine M_{g+d} on pair (x_t, y_t)

Fig. 1 Online learning protocol for post-editing workflow

3 Online adaptation

Online adaptation for NMT follows the online learning protocol shown in Figure 1 that we adopted from Bertoldi et al (2014). The protocol assumes a global model M_g that was trained on a large dataset of parallel data. This dataset does not necessarily come from the same domain as the data used in online adaptation. Online adaptation proceeds by performing online fine-tuning on a further set of patent documents d, resulting in a combined model M_{g+d}. This is done by invoking a sequence of $|d|$ interactions, where on each step a translation output $\hat{y}_t$ for an input source segment x_t is produced by the NMT system, a post-edit y_t is produced by the user, and a system update is performed by using the pair (x_t, y_t) as supervision signal in online learning.

4 Experimental Setup

4.1 NMT System

The NMT system used in our experiments is based on the Lamtram toolkit (Neubig, 2015), which is built on the dynamic neural network library DyNet (Neubig et al, 2017). It implements an encoder-decoder architecture with attention mechanism. The settings in our experiments use dot product as attention type, together with attention feeding (Luong et al, 2015) where the context vector of the previous state is used as input to the decoder neural network. We trained recurrent neural networks (RNNs) with 2 layers consisting of 256 units, and a word representation layer of 128 units, on GPU. The chosen RNN architecture is a long short-term memory network (Hochreiter and Schmidhuber, 1997). As stochastic optimization method we used ADAM (Kingma and Ba, 2015) with a learning rate initialization to 0.001. To prevent overfitting, we set the dropout rate to 0.5 (Srivastava et al, 2014), and used a development set of 2k sentences for early stopping. Evaluation was performed after every 50k sentences.

In order to use Lamtram as an interactive online learning platform, we needed to modify the tool to allow training and translation to take turns without having to reload the model parameters. Furthermore, we implemented a user interface that sends inputs via the network to a web client that renders the source and the proposed NMT for user post-editing, and records post-edits

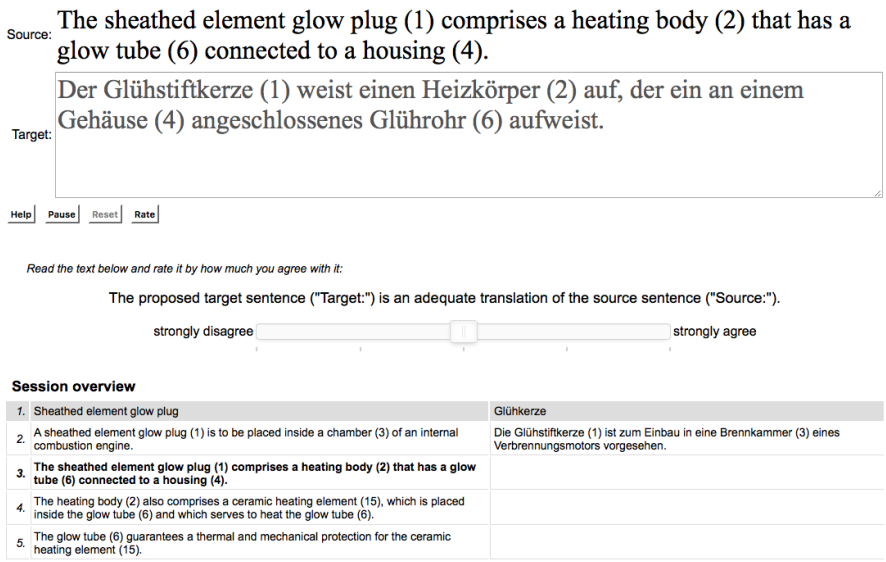

Fig. 2 User interface for post-editing

for learning. A screenshot of the interface is shown in Figure 2. From top to bottom, it shows the source, the post-editing field, and the slider used to collect human quality ratings of the NMT outputs before post-editing. Segments are complete patent abstracts in their original order preluded by the respective patent's title.

For online adaptation we used stochastic gradient descent, with a learning rate of 0.05 and a dropout of 0.25. For inference, the beam size was set to 10, and we tuned a word penalty parameter to adjust the lengths of the outputs, as well as a penalty for unknown words. These parameters were set to 0.85 and 0.25 respectively.

4.2 Data

As training data we used $\sim$ 2M parallel sentences extracted from Europarl and News Commentary. Furthermore, offline fine-tuning was performed on $\sim$ 350k parallel sentences of in-domain data from PatTR[2]. The translation direction is from English into German.

Since patent claims and descriptions tend to be extremely complex and long, they are not suitable for translation by non-experts. We therefore used titles and abstracts for both training and test. Development and test data are limited to documents, each consisting of a patent title and abstract, with an overall maximum length of 45 tokens per sentence. The data split was done by year and by family id to avoid any possible overlaps. The test data were

[2] `http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/`

automatically grouped into clusters by cosine similarity of their bag-of-words tf-idf source representations and length, to obtain clusters of related documents with an approximate source token count of 500, which is appropriate in a post-editing setup given the available time limit of 90 minutes. This way, each cluster contained the titles and complete abstracts of 3-5 documents. All data were preprocessed by tokenization, truecasing, and byte pair encoding (Sennrich et al, 2016b) with a vocabulary size of 10k for source and target, respectively.

4.3 Experimental Design

Our post-editors were 29 master-level students at the Institute for Translation and Interpreting at Heidelberg University. The experiments were conducted during 8 post-editing sessions with a duration of 60-90 minutes each over the course of 5 days. All sessions took place in the same computer pool with the same hardware on each computer. Since the parallel patent resource PatTR is indexed by various sites and can be retrieved through concordance search, in order to receive human post-edits instead of copies of the parallel data, we needed to block web access with the exception of Wikipedia[3] and two online dictionaries that do not crawl for parallel data[4].

In our experiments, we processed overall 105 documents, resulting in 4,563 sentences. The same documents were used to test the effect of the adaptation condition on human post-editing, however, in difference to Bentivogli et al (2016a), no document was seen by the same translator twice. Instead, all documents were translated under two different translation conditions (with and without online adaptation) and post-edited by two different students.[5] In total we collected 4,563 per-sentence measures (NMT adaptation false - 2,354 and true - 2,209) of (h)BLEU (Papineni et al, 2002), (h)TER (Snover et al, 2006), translation quality rating (Graham et al, 2016), keyboard strokes and mouse clicks (Barrachina et al, 2009), and post-editing time.

5 Analysis and Results

5.1 Statistical Analysis

To analyze the results, we used linear mixed-effects models (LMEMs), implemented in the lme4 package (Bates et al, 2015) in R (R Core Team, 2014). Baayen et al (2008) introduced the usage of LMEMs for the analysis of repeated measurement data, enabling to resolve non-independencies by introducing sources of variation, by-subject and by-item variation, as random effects into the model. The general form of an LMEM can be described as the

[3] `https://de.wikipedia.org`

[4] `http://www.dict.cc,http://dict.leo.org`

[5] A single document was used as an exam in the very last session and translated by all post-editors and without adaptation.

Table 1 Excerpt for the model coefficients for the used fixed effect of online NMT adaptation to the individual intercepts for response variable hBLEU; in the example, the global intercept has a value of 47.19 and the global slope lies at 6.73

Random effect	Individual intercept	Individual slope
sentenceID_15	39.64	10.67
sentenceID_16	49.86	19.52
sentenceID_17	53.12	23.73
sentenceID_18	53.39	11.75
sentenceID_19	66.55	20.39
user_A	45.96	3.06
user_B	53.66	10.63
user_C	51.23	15.01
user_D	37.77	5.06
user_E	53.07	7.28
user_F	54.10	8.97

unconditional distribution of a vector of random effects b, and the conditional distribution of a vector-valued random response variable Y given b, which are both multivariate Gaussian distributions (Bates et al, 2015). In matrix form, the LMEM can be expressed by the following formula:

$$Y = X\beta + Zb + \epsilon, \tag{1}$$

where β and b are fixed-effects and random-effects vectors, X and Z are fixed-effects and random-effects design matrices, and ϵ is a vector of random errors.

In our application, the main fixed effect is the NMT adaptation condition (online adaptation to post-edits versus offline learning of a global model only), for which several response variables were measured. The random effects have differing intercepts as well as differing slopes. The granularity of the model is at the sentence level. The observed response value for the i-th subject and j-th sentence, $y_{ij} \in Y$ (for example, a time measurement, hTER, hBLEU, etc.), is defined in our application of LMEMs as follows:

$$y_{ij} = \beta_0 + (\beta_1 + \beta_2)x_{ij} + b_{0i} + b_{0j} + (b_{1i} + b_{1j})z_{ij} + \epsilon_{ij}. \tag{2}$$

The LMEM yields estimates of a global intercept β_0 (i.e., the expected mean value of a response variable when all slopes are equal to 0) and global slopes for the used fixed effects of NMT adaptation β_1 and dayID β_2. Thus in our experiments, a global intercept β_0 is an estimate for an average value for the measurements across all students and all sentences on the first day in the scenario without NMT adaptation. A global slope of the main fixed effect β_1 provides an estimate for the difference due to online NMT adaptation. The global slope β_2 of the dayID fixed effect estimates differences in response variables in consecutive sessions. Furthermore, we get random effect intercepts for subject b_{0i} and sentence levels b_{0j} (i.e., for each level we get that level's intercept's deviation from the global intercept) and random effect slopes within

Table 2 Slopes for the used fixed effects of online NMT adaptation and dayID (index of day when consecutive sessions took place) to the global intercept; the response variable is post-editing time in ms, with a global intercept of 811.76 ± 43.47

Fixed effect	Slope
NMT adaptation	-29.72 ± 27.90
dayID_2	-135.81 ± 38.33
dayID_3	-171.21 ± 38.30
dayID_4	-221.24 ± 38.25
dayID_5	-235.82 ± 35.30

each user b_{1i} and sentence level b_{1j} (i.e., the degree to which a fixed effect deviates from the global slope within a given level). Thus, each student and each sentence get its own individual intercept, or average value, in the scenario without NMT adaptation, as well as its own estimate for improvement due to online NMT adaptation. In our case, $(x_{ij}) = X$ being equal to $(z_{ij}) = Z$, is a design matrix of categorical variables with regard to the measurement for the j-th sentence and i-th subject and the respective predictor; ϵ_{ij} is an error term.

We applied the idea of maximum random effects in our model by using random slopes for each random effect to account for different reactions of subjects and for different effects for items with regard to experimental conditions (Barr et al, 2013). Table 1 illustrates the usefulness of the used maximum random effects structure by a sample of individual intercepts and respective slopes for each level of random effects in the LMEM for hBLEU. First, interpreting individual intercepts as average hBLEU values without NMT adaptation, we observe that both individual intercepts for sentences and post-editors differ from the global intercept, which is the average value of 47.19 found in measurements in the offline learning scenario. The high variance within individual intercepts of each random effect indicates varying difficulty of sentences to be translated and diverse preferences and experience of post-editors. Second, the individual slope values, interpreted as hBLEU improvements due to online NMT adaptation, make evident how much influence the effect of NMT adaptation has on each sentence and for each post-editor. In comparison to a global slope of 6.73, individually estimated random effect slopes vary significantly. This shows that different sentences are harder to improve and different post-editors react differently to manipulations to the NMT system.

Also, we found it useful to add the dayID as an additional fixed effect in the LMEM. The variable dayID labels the days in chronological order when consecutive sessions took place. It indicates the progress of post-editors through time and their improving experience with the NMT system and post-editing practice. An example is given in Table 2 which compares the speedup of the post-editing time in consecutive sessions to the improvement in post-editing time due to NMT adaptation. We can clearly see that the learning effect over

Table 3 Slope for the fixed effect of online NMT adaptation to the global intercept. LMEMs were built for hBLEU, TER/BLEU between NMT output (mt) and reference (ref), TER/BLEU between post-edit (pe) and reference, hTER, rating, count of keystrokes and mouse clicks, KSMR and time as response variables; significance of results was tested with likelihood ratio tests of the full model against the model without the independent variable of interest

Response variable	Intercept	Slope	Significance
hBLEU (%)	47.19 ± 1.20	+6.73 ± 1.01	p <0.001
BLEU mt & ref (%)	20.79 ± 0.97	+1.76 ± 0.27	p <0.001
BLEU pe & ref (%)	22.76 ± 1.22	+1.09 ± 0.63	p <0.1
hTER (%)	35.45 ± 0.82	-4.98 ± 0.72	p <0.001
TER mt & ref (%)	56.28 ± 0.97	-0.95 ± 0.29	p <0.02
TER pe & ref (%)	54.86 ± 1.18	-0.53 ± 0.55	-
rating (0-100)	44.23 ± 2.39	+6.11 ± 1.42	p <0.001
kbd+click (count)	73.15 ± 4.76	-12.13 ± 2.08	p <0.001
KSMR (ratio)	0.52 ± 0.02	-0.07 ± 0.02	p <0.001
time (ms)	811.76 ± 43.47	-29.72 ± 27.90	-

time (shown in the reduction in post-editing time in consecutive sessions) has a larger impact than the effect of online NMT adaptation.

5.2 Experimental Results

Table 3 gives the central results of our analysis: We find significant improvements in post-editing effort due to online adaptation, shown in reduced hTER by nearly 5 points and improved hBLEU up to 6.73 points. Furthermore, online adaptation has a domain adaptation effect which leads to translation outputs which are closer to static references, shown in an increase of BLEU between NMT output and reference translation as well as BLEU between post-edit and reference. This agrees with the quality assessment (rating) that users had to give using a 100-point slider before they can start the post-editing process: Post-editors assess the quality of the NMT output at 6 points higher in case of online adaptation.

In measuring post-editing time, we normalized wall-clock time by the number of characters in a post-edit. The improvement in time is 30 ms less per character, corresponding to a nominal improvement of 3.7%. We conjecture that the reason why we could not establish a significant improvement in post-editing time could be due to improved post-editors' experience in consecutive sessions. Moreover, despite the used by-item and by-subject random intercepts and slopes, we see a high variability of post-editing speed across items and subjects, which makes it difficult to prove significance for the effect of online NMT adaptation for the reduction of post-editing time. This result confirms similar findings reported in Bentivogli et al (2016a). However, as is shown in the

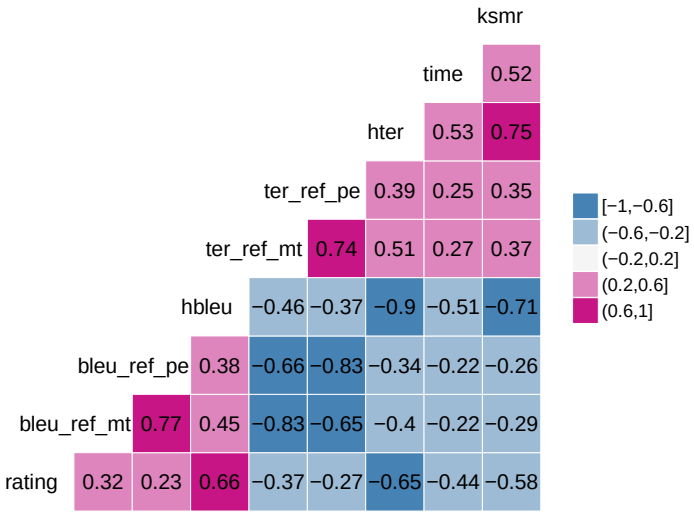

Fig. 3 Correlation matrix for pairwise Pearson correlation coefficients for KSMR, time, hTER, hBLEU, TER/BLEU between reference and post-edit, TER/BLEU between reference and NMT output, rating; strong positive correlations are marked with dark red; strong negative correlations by dark blue

next section, our correlation analysis allows to establish a strong tie between post-editing time and metrics of post-editing effort.

In order to measure technical effort of post-editing, we combined keyboard strokes and mouse clicks into the metric of key-stroke and mouse-action ratio (KSMR) - a measure proposed by Barrachina et al (2009). KSMR is calculated as the sum of the number of keystrokes and the number of mouse movements plus one, divided by the count of characters in the reference. We observed significant reduction in KSMR. According to the LMEM analysis, online NMT adaptation enables post-editors to use 12 keystrokes and mouse clicks less per sentence.

Table 4 shows example patents which were post-edited during our experiment. In the first example, our adapted NMT system has learned the right translation of *blades* after a post-editor changed *Schaufeln* to *Klingen*. In the second example, the technical term *image recorder* was inconsistently and wrong translated by the baseline system as *Bildaufzeichner* or *Bildaufnah-*

Table 4 Examples for test patent data: source, reference, NMT output, post-edit, and adapted NMT output

Source	The outer surfaces of the **blades** (172) are inclined relative to the axis of rotation. When the product to be cut is pushed into the knife arrangement (170), the latter is rotated in such a manner that the **blades** (172) cut the product to be cut along helical paths.
Reference	Die Aussenflächen der Klingen (172) sind relativ zur Drehachse geneigt. Wenn das Schneidgut in die Messeranordnung (170) eingeschoben wird, wird diese derart in Drehung versetzt, dass die Klingen (172) das Schneidgut entlang helikaler Bahnen zerteilen.
NMT output	Die Aussenflächen der *Schaufeln* (172) sind zur Drehrichtungsachse geneigt. Beim Eindringen des zu durchtrennenden Produkts in die Messeranordnung (170) wird das zu schneidende Gut so gedreht.
Post-edit	Die Außenflächen der **Klingen** (172) sind bezogen auf die Rotationsachse neigbar. Beim Einführen des zu schneidenden Produktes in die Messergruppe (170) wird letztere so gedreht, dass die Klingen (172) das zu schneidende Produkt spiralförmig schneiden.
Adapted NMT	Die Aussenflächen der *Schaufeln* (172) sind bezogen auf die Rotationsachse neigbar. Beim Einschnitt des Warengutes in die Messeranordnung (170) wird die **Klinge** so gedreht, dass die **Klinge** (172) das zu schneidende Gut schneidet.
Source	Miniaturized **image recorder** The aim of the invention is to create an **image recorder** which can be miniaturized to such an extent that an endoscope of the smallest dimensions can be produced therewith. The disclosed **image recorder** can also be optimized in such a manner that the major part of the semiconductor surface is made available for pixel integration.
Reference	Miniaturisierter Bildaufnehmer Ziel der offen gelegten Erfindung ist es einen Bildaufnehmer vorzuschlagen, der soweit miniaturisiert werden kann, dass damit ein Endoskop kleinster Dimensionen realisiert werden kann. Ausserdem kann der offen gelegte Bildaufnehmer dahingehend optimiert werden, dass der grösste Teil der Halbleiterfläche für die Integration von Bildpunkten zur Verfügung steht.
NMT output	Miniaturisierter *Bildaufzeichner* Der Erfindung liegt die Aufgabe zugrunde, einen *Bildaufnahmeer* zu schaffen, der so miniaturiert ist, dass eine Impedanz der kleinsten Abmessungen erzielbar ist. Der *Bildaufzeichner* kann auch so optimiert werden, dass der grösste Teil der Halbleiterfläche für die galvanische Integration bereitgestellt wird.
Post-edit	Miniaturisierter **Bildaufnahmeapparat** Das Ziel der Erfindung ist es, einen Bildaufnahmeapparat herzustellen, der so miniaturisiert ist, dass ein Endoskop der kleinsten Abmessungen damit hergestellt werden kann. Der verlautbarte Bildaufnahmeapparat kann auch so optimiert werden, dass der grösste Teil der Halbleiterfläche für die Pixelintegration bereitgestellt wird.
Adapted NMT	Miniaturisierter *Bildlautsprecher* Der Erfindung liegt die Aufgabe zugrunde, einen **Bildaufnahmeapparat** zu schaffen, der so miniaturiert ist, dass eine Impedanz der kleinsten Abmessungen hergestellt werden kann. Der **Bildaufnahmeapparat** kann auch so optimiert werden, dass der grösste Teil der Halbleiterfläche für die galvanische Integration bereitgestellt wird.

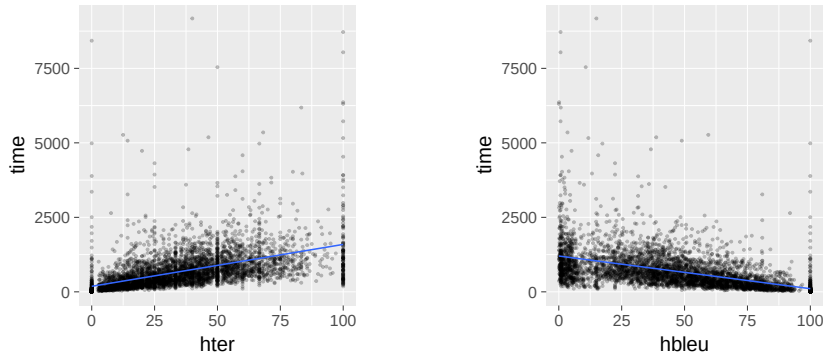

Fig. 4 Regression plots of hTER and time (left) and hBLEU and time (right)

meer. The adapted system learns the translation *Bildaufnahmeapparat* from the post-edit.

5.3 Correlation Study

Since due to the relatively small size of our user study, significance of the result difference between the adaptation conditions could not be established for some response variables, we furthermore analyzed the correlations between the obtained measures. For this purpose, we calculated pairwise Pearson correlation coefficients (Figure 3). The correlation matrix shows strong correlations of translation quality rating with hBLEU, hTER, time and KSMR. Figure 4 shows detailed values for post-editing time plotted against hTER and hBLEU. The plots confirm that while the effect of online adaptation was not significant for the response variable of time, there is a strong correlation of post-editing time with metrics for post-editing effort such as hTER or hBLEU. Figure 5 plots TER and BLEU between NMT and post-edit to human references. While the increase in BLEU and TER between post-edit and reference was barely significant in itself, these plots confirm a strong correlation of improved translation quality of NMT outputs with improved quality of post-edits.

6 Summary and Conclusion

We presented a user-study on the effect of online adaptation on NMT systems to interactive user post-edits of the proposed translations. We found significant reductions in human post-editing effort along several well-established response variables (hTER, hBLEU, KSMR).

Furthermore, we found a domain adaptation effect due to online adaptation, leading to significant improvements of TER/BLEU of the machine translations with respect to human reference translations. Since acquiring a

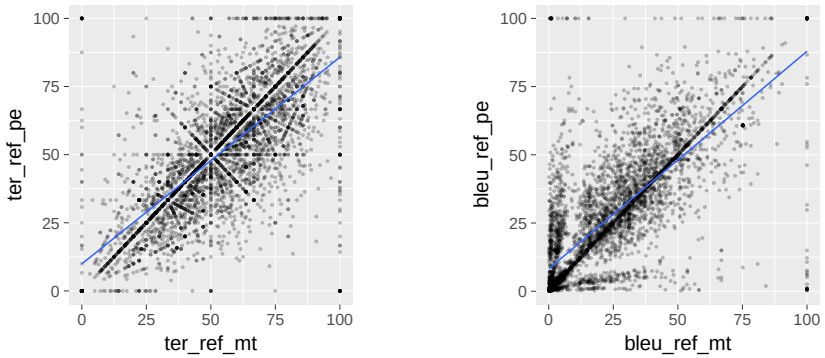

Fig. 5 Regression plots of TER between NMT output and reference and TER between post-edit and reference (left) and BLEU between NMT output and reference and BLEU between post-edit and reference (right)

domain-specific vocabulary as indicated by improved TER/BLEU is an important quality metric in the domain of patent translation, we also measured and established a strong correlation of the machine translation improvements with TER/BLEU between translation students' post-edits and professional patent translators' references. This shows a beneficial effect of improved quality of machine translations on improved quality of post-edits.

Due to our experimental setup where the same documents were translated by both a static and an adaptive NMT system, and post-edited by two different translators at different points in time, we found a confounding effect between improved post-editing experience and reduced time. This did not allow us to establish significant improvements of online NMT adaptation with respect to post-editing time. However, we found a strong correlation of reduction in post-editing time to improvements in metrics for post-editing effort such as hTER, hBLEU, or KSMR. This shows firstly that the latter metrics are more reliable indicators for reduced post-editing effort, and furthermore that reduced post-editing time is a correlated, but not necessarily directly causally related effect.

In sum, our user study established significant improvements due to online NMT adaptation along well-known metrics of post-editing effort, and along the dimension of domain adaptation that is particularly important in technical translation domains such as patent translation. Our study did not touch novel modes of user interaction with NMT systems (for example, human bandit feedback: Kreutzer et al, 2017; Nguyen et al, 2017) or alternative modes of NMT system adaptation (for example, interactive translation prediction: Knowles and Koehn, 2016). These topics are subject of future work.

Acknowledgements The research reported in this paper was supported in part by the German research foundation (DFG) under grant RI-2221/2-1.

References

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language 59(4):390–412

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA

Barr DJ, Levy R, Scheepers C, Tilly HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68(3):255–278

Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lagarda A, Ney H, Tomás J, Vidal E, et al (2009) Statistical approaches to computer-assisted translation. Computational Linguistics 35(1):3–28

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48

Bentivogli L, Bertoldi N, Cettolo M, Federico M, Negri M, Turchi M (2016a) On the evaluation of adaptive machine translation for human post-editing. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24(2):388–399

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016b) Neural *versus* phrase-based machine translation quality: a case study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX

Bentivogli L, Bisazza A, Cettolo M, Federico M (2018) Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. Computer Speech & Language 49:52–70

Bertoldi N, Simianer P, Cettolo M, Wäschle K, Federico M, Riezler S (2014) Online adaptation to post-edits for phrase-based statistical machine translation. Machine Translation 29:309–339

Burchardt A, Macketanz V, Dehdari J, Heigold G, Peter JT, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. The Prague Bulletin of Mathematical Linguistics 108(1):159–170

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017a) Is neural machine translation the new state of the art? The Prague Bulletin of Mathematical Linguistics 108(1):109–120

Castilho S, Moorkens J, Gaspari F, Sennrich R, Sosoni V, Georgakopoulou Y, Lohar P, Way A, Miceli Barone A, Gialama M (2017b) A comparative quality evaluation of PBSMT and NMT using professional translators. In: Proceedings of MT Summit: Research Track

Cesa-Bianchi N, Reverberi G, Szedmak S (2008) Online learning algorithms for computer-assisted translation. Tech. rep., SMART

Denkowski M, Dyer C, Lavie A (2014a) Learning from post-editing: Online model adaptation for statistical machine translation. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden

Denkowski M, Lavie A, Lacruz I, Dyer C (2014b) Real time adaptive machine translation for post-editing with cdec and transcenter. In: Proceedings of the EACL Workshop on Humans and Computer-assisted Translation, Gothenburg, Sweden

Farajian MA, Turchi M, Negri M, Bertoldi N, Federico M (2017) Neural vs. phrase-based machine translation in a multi-domain scenario. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)

Forcada ML (2017) Making sense of neural machine translation. Translation Spaces 6(2):291–309

Graham Y, Baldwin T, Moffat A, Zobel J (2016) Can machine translation systems be evaluated by the crowd alone? Natural Language Engineering 23(1):3–30

Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM

Green S, Wang S, Chuang J, Heer J, Schuster S, Manning CD (2014) Human effort and machine learnability in computer aided translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar

Hardt D, Elming J (2010) Incremental re-training for post-editing SMT. In: Proceedings of the Conference of the Association for Machine Tranlation in the Americas (AMTA), Denver, CO

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780

Isabelle P, Cherry C, Foster G (2017) A challenge set approach to evaluating machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)

Jean S, Firat O, Cho K, Memisevic R, Bengio Y (2015) Montreal neural machine translation systems for WMT'15. In: Proceedings of the Workshop on Statistical Machine Translation (WMT), Lisbon, Portugal

Junczys-Dowmunt M, Dwojak T, Hoang H (2016) Is neural machine translation ready for deployment? A case study on 30 translation directions. CoRR abs/1610.01108

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. The Prague Bulletin of Mathematical Linguistics 108(1):121–132

Klubička F, Toral A, Sánchez-Cartagena VM (2018) Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. Machine Translation pp 1–21

Knowles R, Koehn P (2016) Neural interactive translation prediction. In: Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Austin, TX

Koehn P, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation

Kreutzer J, Sokolov A, Riezler S (2017) Bandit structured prediction for neural sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada

López-Salcedo FJ, Sanchis-Trilles G, Casacuberta F (2012) Online learning of log-linear weights in interactive machine translation. In: Proceedings of IberSpeech, Madrid, Spain

Luong M, Manning CD (2015) Stanford neural machine translation systems for spoken language domains. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Da Nang, Vietnam

Luong M, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal

Macketanz V, Avramidis E, Burchardt A, Helcl J, Srivastava A (2017) Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. Cybernetics and Information Technologies 17(2):28–43

Martínez-Gómez P, Sanchis-Trilles G, Casacuberta F (2012) Online adaptation strategies for statistical machine translation in post-editing scenarios. Pattern Recognition 45(9):3193–3202

Menacer MA, Langlois D, Mella O, Fohr D, Jouvet D, Smaïli K (2017) Is statistical machine translation approach dead? In: Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)

Nakov P, Guzman F, Vogel S (2012) Optimizing for sentence-level BLEU+1 yields short translations. In: Proceedings of the Conference on Computational Linguistics (COLING), Mumbai, India

Nepveu L, Lapalme G, Langlais P, Foster G (2004) Adaptive language and translation models for interactive machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain

Neubig G (2015) lamtram: A toolkit for language and translation modeling using neural networks. http://www.github.com/neubig/lamtram

Neubig G, Dyer C, Goldberg Y, Matthews A, Ammar W, Anastasopoulos A, Ballesteros M, Chiang D, Clothiaux D, Cohn T, Duh K, Faruqui M, Gan C, Garrette D, Ji Y, Kong L, Kuncoro A, Kumar G, Malaviya C, Michel P, Oda Y, Richardson M, Saphra N, Swayamdipta S, Yin P (2017) Dynet: The dynamic neural network toolkit. CoRR abs/1701.03980

Nguyen K, Daumé H, Boyd-Graber J (2017) Reinforcement learning for bandit neural machine translation with simulated feedback. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark

Ortiz-Martínez D, García-Varea I, Casacuberta F (2010) Online learning for interactive statistical machine translation. In: Proceedings of the Human Language Technologies conference and the Annual Conference of the North

American Chapter of the Association for Computational Linguistics (HLT-NAACL), Los Angeles, CA

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), Stroudsburg, PA

Peris Á, Domingo M, Casacuberta F (2017) Interactive neural machine translation. Computer Speech & Language 45:201–220

Popović M (2017) Comparing language related issues for NMT and PBMT between german and english. The Prague Bulletin of Mathematical Linguistics 108(1):209–220

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Sennrich R, Haddow B, Birch A (2016a) Edinburgh Neural Machine Translation Systems for WMT'16. In: Proceedings of the Conference on Machine Translation (WMT), Berlin, Germany

Sennrich R, Haddow B, Birch A (2016b) Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany

Shterionov D, Casanellas PNL, Superbo R, O'Dowd T (2017) Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In: Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT): User Track

Simianer P, Karimova S, Riezler S (2016) A post-editing interface for immediate adaptation in statistical machine translation. In: Proceedings of the Conference on Computational Linguistics: System Demonstrations (COLING Demos), Osaka, Japan

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Cambridge, MA

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)

Turchi M, Negri M, Farajian MA, Federico M (2017) Continuous learning from human post-edits for neural machine translation. The Prague Bulletin of Mathematical Linguistics 108(1):233–244

Wuebker J, Green S, DeNero J, Hasan S, Luong M (2016) Models and inference for prefix-constrained machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany