A User Study of the Incremental Learning in NMT

Miguel Domingo[1] and **Mercedes García-Martínez**[2] and **Álvaro Peris**[3] and **Alexandre Helle**[2] and **Amando Estela**[2] and **Laurent Bié**[2] and **Francisco Casacuberta**[1] and **Manuel Herranz**[2]
[1]PRHLT Research Center - Universitat Politècnica de València
{midobal, fcn}@prhlt.upv.es
[2]Pangeanic / B.I Europa - PangeaMT Technologies Division
{m.garcia, a.helle, a.estela, l.bie, m.herranz}@pangeanic.com
[3]Independent Researcher
lvapeab@gmail.com

Abstract

In the translation industry, human experts usually supervise and post-edit machine translation hypotheses. Adaptive neural machine translation systems, able to incrementally update the underlying models under an online learning regime, have been proven to be useful to improve the efficiency of this workflow. However, this incremental adaptation is somewhat unstable, and it may lead to undesirable side effects. One of them is the sporadic appearance of made-up words, as a byproduct of an erroneous application of subword segmentation techniques. In this work, we extend previous studies on on-the-fly adaptation of neural machine translation systems. We perform a user study involving professional, experienced post-editors, delving deeper on the aforementioned problems. Results show that adaptive systems were able to learn how to generate the correct translation for task-specific terms, resulting in an improvement of the user's productivity. We also observed a close similitude, in terms of morphology, between made-up words and the words that were expected.

1 Introduction

Despite its improvements and obtaining admissible results in many tasks, machine translation (MT) is still very far from obtaining automatic high-quality translations (Dale, 2016; Toral et al., 2018). Thus, a human agent needs to supervise and correct the outputs generated by an MT system. This process is known as *post-editing* and is a common use case of MT in the industrial environment. As MT systems are continuously improving their capabilities, it has acquired major relevance in the translation market (Guerberof, 2008; Pym et al., 2012; Hu and Cadwell, 2016; Turovsky, 2016).

Throughout the post-editing process, new data are continuously generated. These new data have valuable properties—they are domain-specific training samples. Thus, it can be leveraged to continuously adapt the system towards a given domain or the style of the post-editor. A common way of achieving this consists in following an online-learning paradigm (Ortiz-Martínez, 2016; Peris and Casacuberta, 2019). Each time the user validates a post-edit, the system's models are updated incrementally with this new sample. Hence, when the system generates the next translation, it will consider the previous post-edits made by the user and it is expected to produce higher quality translations (or, at least, more suited to the post-editor's preferences).

Domingo et al. (2019b) conducted a preliminary user study for professional post-editors, who had a positive perception of the adaptive systems. However, they noticed that, in some cases, there were occurrences of some made-up words. In this work, we study the impact of this phenomenon. Additionally, we extend their user study by involving three more participants and providing additional measures for the increase in productivity gained with the adaptive system.

2 Related work

Post-editing MT hypotheses is a practice that was adopted in the translation industry a long time ago (e.g., Vasconcellos and León, 1985). Its relevance grew as MT technology advanced and improved. The capabilities of MT post-editing have been demonstrated through many user studies (Aziz et al., 2012; Bentivogli et al., 2016; Castilho et al., 2017; Green et al., 2013a).

Parallel to the rise of the post-editing protocol, using user post-edits to adapt MT systems has also attracted the attention of researches and industry. This was studied in the CasMaCAT (Alabau et al., 2013) and MateCAT (Federico et al., 2014) projects and phrase-based statistical MT systems based on online learning were developed (Ortiz-Martínez, 2016). With the breakthrough in neural MT (NMT) technology (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017), research shifted towards constructing adaptive systems via online learning in this post-editing scenario. The use of online learning to adapt an NMT system to a new domain with post-edited samples was proposed by Peris et al. (2017) and Turchi et al. (2017). Other works refined these adaptation techniques and applied them to new use cases (Kothur et al., 2018; Wuebker et al., 2018; Peris and Casacuberta, 2019).

The evaluation of MT post-edits is a hard topic that

Corpus	#Sentences	# Tokens		# Types		Average length	
		En	Es	En	Es	En	Es
Training	23.4M	702M	786M	1.8M	1.9M	30.0	33.6
Document 1	150	1.7K	-	618	-	11.3	-
Document 2	150	2.6K	-	752	-	17.3	-

Table 1: Corpora statistics in terms of number of sentences, number of tokens, number of types (vocabulary size) and average sentence length. K denotes thousands and M, millions.

is currently being actively researched (e.g., Toral, 2019; Freitag et al., 2020; Läubli et al., 2020). Several works conducted user studies for MT post-editing systems, either phrase-based (Alabau et al., 2013; Green et al., 2013b; Denkowski et al., 2014; Bentivogli et al., 2016) or NMT (Daems and Macken, 2019; Koponen et al., 2019; Jia et al., 2019). Moreover, two studies showed improvements in terms of productivity time and translation quality with the application of an online learning protocol (Karimova et al., 2018; Domingo et al., 2019b). This latter study is tightly related to ours. We extend it by performing a finer-grained evaluation of the outputs of the adaptive systems.

3 Experimental framework

As we extended the work of Domingo et al. (2019b), we used their same data and systems. The task at hand consisted of a small medico-technical (description of medical equipment) corpus from their production scenario. It contains specific vocabulary from a very closed domain. It was conformed by two documents of 150 sentences, which contained 1.7 and 2.7 thousand words respectively. The translation direction was English to Spanish. The system was trained using the data from WMT'13's translation task (Bojar et al., 2013) and samples selected by the feature decay selection technique (Biçici and Yuret, 2015). The data features are summarized in Table 1. We applied joint byte pair encoding (Sennrich et al., 2016), using $32,000$ merge operations. The system was built with OpenNMT-py (Klein et al., 2017), using a long short-term memory (Gers et al., 2000) recurrent encoder–decoder with attention (Bahdanau et al., 2015). All model dimensions were 512. The system was trained using Adam (Kingma and Ba, 2014) with a fixed learning rate of 0.0002 (Wu et al., 2016) and a batch size of 60. A label smoothing of 0.1 (Szegedy et al., 2015) was applied. At inference time, we used beam search with size 6.

The adaptation process followed the findings from Peris and Casacuberta (2019). We tuned the hyperparameters for the adaptation process on our development set, under simulated conditions. For each new post-edited sample, we applied two plain SGD updates, with a fixed learning rate of 0.05.

As translation environment we used the one designed by Domingo et al. (2019a). It connects our adaptive NMT engine with the SDL Trados Studio interface, which is used by the post-editors in our production workflow. In addition, it also allowed us to trace the productivity metrics and user behavior.

3.1 Evaluation

We evaluated two main aspects of the adaptation process: productivity of the post-editors and quality of the NMT systems. The former was assessed by computing the average post-editing time per sentence and the number of words generated by the post-editor per hour. For the latter, we employed two well-known MT metrics: (h)BLEU (Papineni et al., 2002) and (h)TER (Snover et al., 2006). In order to ensure consistent BLEU scores, we used sacreBLEU (Post, 2018). Since we computed per-sentence BLEU scores, we used exponential BLEU smoothing (Chen and Cherry, 2014).

We applied approximate randomization tests (Riezler and Maxwell, 2005), with $10,000$ repetitions and a p-value of 0.05, to determine whether two systems presented statistically significant differences.

3.2 Human post-editors

Six professional translators were involved in the experiment. Some profiling details about them can be found in Table 2.

User	Sex	Age	Professional experience
User 1	Male	24	1.5 years
User 2	Female	25	5 years
User 3	Female	30	5 years
User 4	Female	24	1 month
User 5	Female	22	1 year
User 6	Male	48	22 years

Table 2: Information about the human post-editors that took part in the experiment, regarding their sex, age and years of professional experience.

The static experiment consisted in post-editing using the initial NMT system, which remained fixed along the complete process. For the adaptive experiment, all users started with the initial system, which was adapted to each user through the process using their own post-edits. Therefore, at the end of the process, each user obtained a tailored system. In order to avoid the influence of translating the same text multiple times, each participant post-edited a different document set under each scenario (static and adaptive), as shown in Table 3.

User	Document 1	Document 2
User 1	Static	Adaptive
User 2	Adaptive	Static
User 3	Static	Adaptive
User 4	Adaptive	Static
User 5	Static	Adaptive
User 6	Adaptive	Static

Table 3: Distribution of users, document sets and scenarios. All users conducted first the experiment which involved post-editing document 1 and then document 2 (e.g., user 2 first post-edited document 1 on an adaptive scenario and, then, document 2 on a static scenario).

4 User study

In our study, we focus on the differences between static and adaptive systems based on three main aspects: the productivity of post-editors, the quality of post-edits and the generation differences.

4.1 On the productivity of the post-editors

Table 4 shows the average gains obtained in terms of translation quality. These results demonstrate how the adaptive systems benefits from the user post-edits to improve the translation quality, yielding gains of up to 6.7 TER points and 8.0 BLEU points.

Test	System	hTER [↓]	hBLEU [↑]
Document 1	Static	39.5	47.9
	Adaptive	32.8†	55.9†
Document 2	Static	36.2	42.9
	Adaptive	34.3†	50.5†

Table 4: Results of the user experiments, in terms of translation quality. These numbers are averages over the results obtained by the different post-editors. *Static system* stands for conventional post-editing—without adaptation. *Adaptive system* refers to post-editing in an environment with online learning. *hTER* and *hBLEU* refer to the TER and BLEU of the system hypothesis computed against the post-edited sentences. † indicates statistically significant differences between the static and the adaptive systems.

Table 5 presents the productivity improvements yielded by the adaptive system. With two exceptions, the adaptive system significantly reduced the averaged time needed to post-edit a sentence (with gains from 4.0 up to 33.5 seconds per sentence). These two exceptions were for user 2—whose average time was the same for both systems—and user 4—whose average time was bigger when using the adaptive system. This last case can be explained by taking into account that user 4 is one of the least experienced users and that she conducted the experiment involving the adaptive scenario first (see Tables 2 and 3). Therefore, as time goes on, user 4 feels more comfortable with the task and tools and, thus, the post-editing time decreases. This phenomenon was already observed during the CasMaCAT project (Alabau et al., 2013).

When measuring productivity in terms of number of words generated per hour, the adaptive systems achieved significant gains for all cases except for user 4—which is coherent with the results obtained in terms of time per sentence. These gains range from 6—for user 2, who took the same average time for both scenarios—to 686 words per hour. Therefore, both metrics showcase how adaptive systems are able to significantly improve productivity.

User	System	Time [↓]	Words per hour [↑]
User 1	Static	37.9	1685
	Adaptive	33.0†	1935†
User 2	Static	30.5	2091
	Adaptive	30.4	2097†
User 3	Static	38.0	1678
	Adaptive	27.0†	2364†
User 4	Static	37.5	1701
	Adaptive	47.4†	1346†
User 5	Static	80.2	795
	Adaptive	46.7†	1367†
User 6	Static	53.7	1188
	Adaptive	49.7†	1284†

Table 5: Results of the user experiments, in terms of productivity. *Static system* stands for conventional post-editing, without adaptation. *Adaptive system* refers to post-editing in an environment with online learning. *Time* corresponds to the average post-editing time per sentence, in seconds. *Words per hour* represents the number of words generated by the post-editors per hour. Users 4 to 6 has less experience, in this particular domain, than users 1 to 3. † indicates statistically significant differences between the static and the adaptive systems.

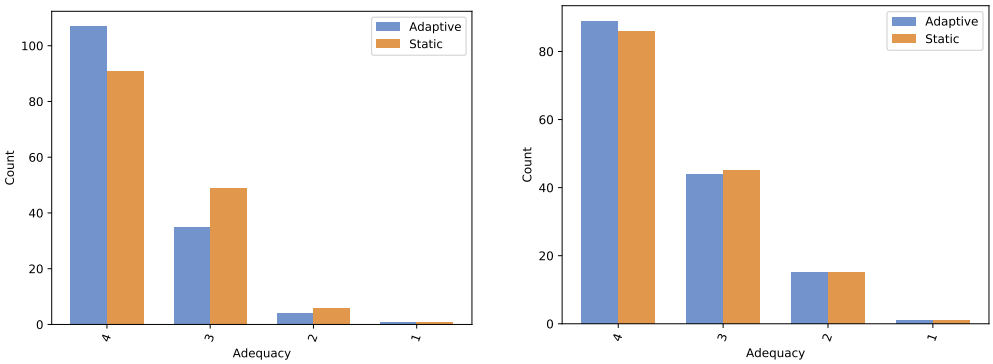
4.1.1 User feedback

Following Domingo et al. (2019b) post-editors filled a questionnaire (see Appendix A) regarding the task they had just performed. We asked them about their level of satisfaction of the translations they had produced; whether they would have preferred translating from scratch instead of post-editing; and their opinion about the automatic translations, in terms of grammar, style and overall quality. Additionally, we also requested them to give, as an open-answer question, their feedback on the task.

While post-editors were generally satisfied with the system and the translations they produced (as also reported by Domingo et al. (2019b)), they spotted some issues regarding the adaptive NMT system: they noticed that domain-specific term were "forgotten" by the system, being wrongly translated. In addition, the users spotted the occurrence of some nonexistent words in the target language (e.g., "absolvido"). We delve deeper into these problems in Section 5.

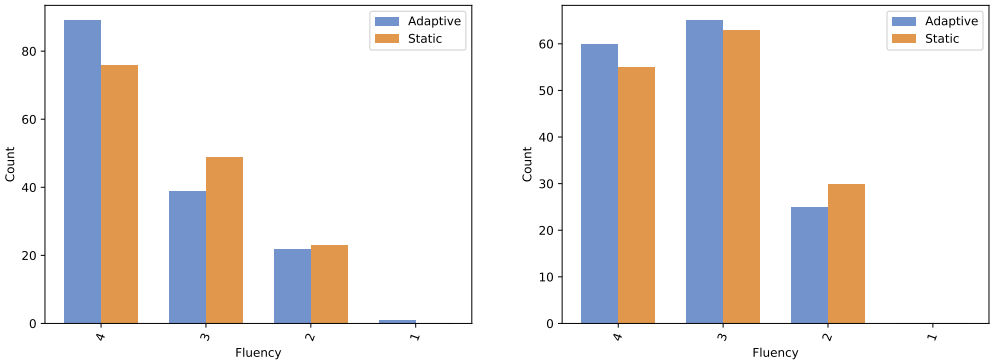
4.2 On the quality of the post-edits

In order to assess and compare the quality of the human post-edits using the static and adaptive systems, a



(a) Document 1. **(b)** Document 2.

Figure 1: Sentence-level adequacy scores. Count values are the average between both evaluators.



(a) Document 1. **(b)** Document 2.

Figure 2: Sentence-level fluency scores. Count values are the average between both evaluators.

human evaluation was conducted with the help of two professional translators—who had not taken part in the user study. In this evaluation, the evaluators were given a source sentence and the post-edits produced by each user—three of which had used the static system, and the other three the adaptive system.

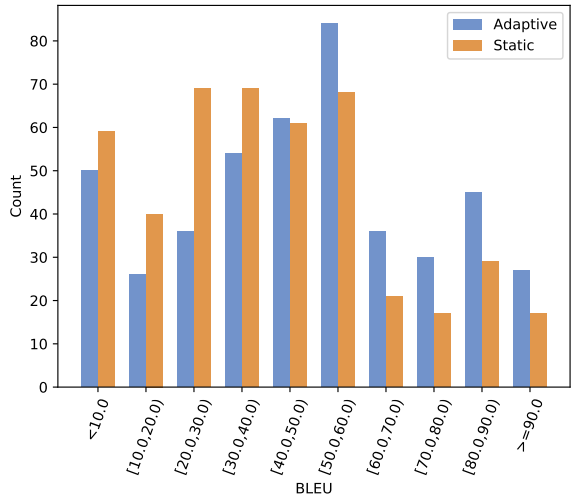
Following Castilho et al. (2019) and TAUS adequacy/fluency guidelines[1], they were asked to assess, on a 4-point scale, the adequacy (how much of the meaning is represented in the translation) and the fluency (the extent to which the translation is well-formed grammatically, has correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker) of each post-edit.

In total, they evaluated 600 sentences: the post-edits of the first 50 sentences of Document 1 and the post-edits from the first 50 sentences of Document 2 (see Section 3). To avoid biases, evaluators were not given any information regarding the origin of the translations. Figs. 1 and 2 present the results of the evaluation.

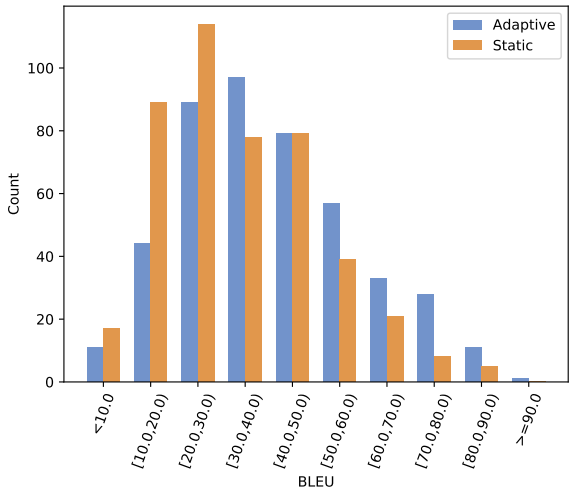
In terms of adequacy, results show that, for both systems, most of the post-edits convey the full meaning of the original sentence or most of it (represented by the scores 4 and 3). Just a few of them convey little or none of the original meaning (represented by the scores of 2 and 1). While both system behave similarly, we observe that a larger amount of the post-edits generated using the adaptive system have the highest adequacy score. This difference is more noteworthy for the post-edits from document 1 than for those from document 2. Similar conclusions can be reached according to fluency: Most post-edits, independently of the system used, are either flawless or good (represented by scores 4 and 3) regarding the extent to which they are constructed. Just a few are considered to be dis-fluent or incomprehensible (represented by a score of 2). Again, both systems are perceived to be similar in document 2, while the adaptive system is perceived as slightly more fluent.

Finally, it is worth noting sine particularities of the task that may have influenced the results of the evaluation: the task consists in the description of medical equipment and, thus, contains several singularities such as specific acronyms (with which the target audience may be more familiar in their original language than with their translation) or description of parts of an equipment (taking into account that the physical equipment may have tags in its original language). Since the evaluators were given no specific instruction about how

[1] https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines.



(a) Document 1.



(b) Document 2.

Figure 3: Histogram of sentence-level BLEU scores. The counts are distributed in buckets of range 10.

to solve those particularities, their personal criteria may had an impact in the evaluation results.

4.3 On differences in the generation

Next, we compare both adaptive and static systems in terms of the translations generated. To this end, we employed the discriminative language model method (Akabe et al., 2014) implemented in the `compare-mt` (Neubig et al., 2019) tool, comparing sentence-level BLEU and word n-grams.

In terms of translation quality, we show a histogram of sentence-level BLEU scores in Fig. 3. For both documents, we observe similar trends: the static system generated low-scored sentences more frequently than the adaptive systems. The adaptive systems placed more hypotheses from bucket $[50, 60)$ onward, for both test documents. Moreover, the differences in frequencies between adaptive and static systems were kept at a similar proportion along all high-score buckets. Hence, adaptive systems were able to outperform the static one in these high-score ranges.

The study of the different n-grams helped us to identify common patterns across all users: adaptive systems were able to effectively learn ad-hoc sequences for the task at hand. We discovered several phenomena among the most common n-gram matches of adaptive systems: the correct translation of acronyms, entities relating a particular device and specific task terminology. See Fig. 4 for examples of these phenomena. We found these common constructions to be one of the major causes of the differences in terms of translation quality.

5 Generation of made-up words

On their feedback, the users reported that, in some cases, the system's hypothesis contained words which were not real words (e.g., "absolvido"). This phenomenon, although infrequent, was a bit cumbersome. Most likely, it is caused by an incorrect segmentation of a word via the byte pair encoding process which, according to their frequency, splits words into multiple tokens. In order to assess its impact, we start by quantifying the issue. Table 6 shows the total of made-up words generated per user.

User	System	Words
User 1	Static	3
	Adaptive	6
User 2	Static	8
	Adaptive	5
User 3	Static	3
	Adaptive	17
User 4	Static	8
	Adaptive	5
User 5	Static	3
	Adaptive	14
User 6	Static	8
	Adaptive	4

Table 6: Total made-up words generated per user.

While this phenomenon is not very frequent (it represents from 0.2 up to 0.8% of all the words generated by a given system), it is present in all systems. Depending on the user, this problem was more present using the static or the adaptive system. Since users were using a different document set for each scenario (see Table 3) and there is a significant difference between documents in terms of total words and vocabulary (see Table 1), we need to compute the average per document in order to evaluate how the problem of made-up words generation affects the different scenarios. These results are shown in Table 7.

Although it could be expected for the adaptive sys-

Phenomenon	System	Example
Acronyms	Source	QSE Number
	Post-edit	Número **de ESC**
	Adaptive	Número **de ESC**
	Static	Número QSE
Entities	Source	Show the R Series ALS
	Post-edit	Mostrar **la serie R** ALS
	Adaptive	Mostrar **la serie R** ALS
	Static	Mostrar el R Series ALS
Terminology	Source	There are several steps involved with sidestream end tidal CO2 setup.
	Post-edit	La configuración del CO2 **espiratorio final** de flujo lateral se realiza en varios pasos.
	Adaptive	Hay varias etapas de la configuración del CO2 **espiratorio final** del ajuste.
	Static	Hay varias etapas que involucran la configuración del CO2 maremoto del CO2 maremoto

Figure 4: Examples of the n-gram differences between adaptive and static systems. In **boldface** we highlight the differences introduced by adaptive systems.

Document	**System**	**Words**
Document 1	Static	5
	Adaptive	4
Document 2	Static	8
	Adaptive	12

Table 7: Average of made-up words generated per document for all users.

tem—which has to deal with a higher number of out-of-vocabularies, introduced by the user—to generate made-up words with a higher frequency, both systems behave similarly: on document 1 case, the static system generated 0.1% more made-up words and, in the other case (document 2), it was the adaptive system which generated 0.1% more made-up words. Furthermore, when comparing both documents, we observe that, despite document 2 having a bigger vocabulary, both static systems generated the same percentage of made-up words. However, Document 2's adaptive systems generated 0.2% of made-up words on average. Most likely, since we did not have an in-domain corpus for training the systems (see Section 3), the bigger the document's vocabulary is, the easier an out-of-vocabulary word may results in an incorrect subword segmentation.

5.1 Error analysis

Fig. 5 shows some example of made-up words generated by the static system.

From the examples, we observe that while the made-up words do not have any sense, they resemble real words (e.g., *pacio* resembles *espacio*; *escaga* resembles *escala*; etc). However, the words they resemble are semantically very different to the correct words (e.g., while *pacio* resembles *espacio*, the correct word would be *estimulación*).

The adaptive systems generates similar made-up words (see Fig. 6 for some examples). However, in this case we observe that some made-up words are almost correct: while *los válvulos* does not exist (*valve* is a feminine word in Spanish), it would be correct, from a morphological point of view, if *valve* were masculine. Something similar, but with the opposite gender, happens with *las marcapasas* (which should be *los marca-*

1. La zona verde es para **pacio**.
2. Roll al paciente a su lado, y luego rodar el electrodo hacia la espalda del paciente a la izquierda de su columna y debajo de la **escaga**.
3. Presione la tecla del **softón**.
4. Sin embargo, el metrónomo **absolvido** si las compresiones son inferiores a las directrices.
5. Que el dispositivo puede hacer un choque de prueba de 30 **jojuelas**.

Figure 5: Example of sentences with made-up words (denoted in **bold**) from the static system. The first word should have been *estimulación*, the second one *omóplato*, the third one *RCP*, the fourth one *sonará* and the fifth one *julios*.

1. Al mover el Selector de modo a Pacer se activará la puerta del **pidante** para abrir.
2. Coloque el sensor con el adaptador instalado fuera de todas las fuentes de CO2 (incluidos los **válvulos** de aire de respiración y respiratorio) exhalado.
3. Las **marcapasas** de estimulación deben producirse aproximadamente cada centímetro en la tira.
4. El conector de autoprueba funciona solo cuando el envase del electrodo es **inabierto** y conectado a la serie R Series.
5. Para aplicar los electrodos OneStep, introduzca primero el electrodo trasero para evitar la **herración** del electrodo delantero.

Figure 6: Example of sentences with made-up words (denoted in **bold**) from the adaptive systems. The first word should have been *marcapasos*, the second one *válvulas*, the third one *marcadores*, the fourth one *cerrado* and the fifth one *deformación*.

pasos) although, in this case, the correct word would be *marcadores*. While we do not have the means for evaluating the impact in the cognitive effort, we believe this kind of errors are more difficult for the users to deal with due to their similarity with the correct words. However, we need to assess the real impact in a future work.

When comparing both type of systems, there are times in which the adaptive systems are able to generate the correct word when the static system had generated a made-up word; times in which the adaptive systems generate the same made-up word than the static system; and times in which the adaptive systems generate a made-up word when the static system was able to generate the correct word. Note that the behavior of the adaptive systems depend on their user (see Fig. 7 for an example).

Static system: Coloque el sensor con el adaptador instalado fuera de todas las fuentes de CO2 (incluido el del paciente) y ***sus válvulas*** de escape para el aire libre exhalado y el ventilador del ventilador.

Adaptive system$_{\text{User 1}}$**:** Coloque el sensor con el adaptador instalado fuera de todas las fuentes de CO2 (incluido el del paciente y ***su respiración*** y el respirador exhalado) .

Adaptive system$_{\text{User 3}}$**:** Coloque el sensor con el adaptador instalado fuera de todas las fuentes de CO2 (incluidos ***los válvulos*** de aire de respiración y respiratorio) exhalado.

Adaptive system$_{\text{User 5}}$**:** Coloque el sensor con el adaptador alejado de todas las fuentes de CO2 incluido el paciente, y ***sus válvulas*** de respiración y respiración exhalados).

Figure 7: Example of the different behaviors of the adaptive systems. At a certain point of the translation hypothesis, the static system generates the words *sus válvulas*. In their place, the adaptive system for user 1 generates the words *su respiración*. However, the adaptive system for user 3 generates the words *los válvulos*—making-up the word *válvulos*. Finally, the adaptive system for user 5 coincides with the static system in generating the words *sus válvulas*.

Finally, we tried to compare, using edit distance, the made-up words with the closest words (in morphological terms) from the vocabulary in order to have a better understanding of this phenomenon. However, this study did not show any significant information: in almost all the cases, made-up words had a lot of morphological similitudes with words from the vocabulary but those words had no semantic relation with the correct word.

6 Conclusions and future work

In this work, we extended a previous user study of an adaptive NMT system. We conducted new experiments with the help of professional translators, and observed significant improvements of the translation quality—measured in terms of hTER and hBLEU—and significant improvements of the user's productivity—measured in terms of post-editing time and number of words generated. We also conducted, with the help of two additional professional translators, a human evaluation that verified the quality of the post-edits generated during the user study.

The users were pleased with the system. They noticed that corrections applied on a given segment generally were reflected on the successive ones, making the post-editing process more effective and less tedious. When comparing the translations generated by both kind of systems, we identified that adaptive systems were able to generate the correct translation of acronyms, entities relating a particular device and specific task terminology.

An undesirable side effect mentioned by the users was the sporadic apparition of made-up words. We studied this phenomenon and reached the conclusion that due to the increase in the number of out-of-vocabularies as part of the post-editing process, adaptive systems suffer this problem more than static systems. Furthermore, sometimes these made-up words are very similar, in morphological terms, to the correct words—such as a feminine word converted into its non-existent masculine equivalent—which made them harder to detect. However, the cognitive impact in the post-editors will need to be assesed before reaching categorical conclusions.

In regards to future work, we should try to assess the cognitive impact of the made-up words phenomenon. We would also like to study the degradation of domain-specific terms, and analyze the impact on the amount of work required to post-edit subsequent sentences as the user provides corrected examples. Additionally, we will integrate our adaptive systems together with other translation tools, such as translation memories or terminological dictionaries, with the aim of fostering the productivity of the post-editing process. With this feature-rich system, we would like to conduct additional experiments involving more diverse languages and domains, using domain-specialized NMT systems, testing other models (e.g., Transformer, Vaswani et al., 2017) and involving a larger number of professional post-editors. Finally, we also intend to implement the interactive–predictive machine translation protocol (Lam et al., 2018; Peris and Casacuberta, 2019) in our translation environment, and compare it with the regular post-editing process.

Acknowledgements

The authors wish to thank the anonymous reviewers for their careful reading and in-depth criticisms and suggestions. The research leading to these results has received funding from the Spanish Centre for Technological and Industrial Development (Centro para el Desarrollo Tecnológico Industrial) (CDTI); the European Union through *Programa Operativo de Crecimiento Inteligente* (Project IDI-20170964) and through *Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER)* from Comunitat Valenciana

(20142020) under project *Sistemas de frabricación inteligentes para la indústria 4.0* (grant agreement IDIFEDER/2018/025); and Generalitat Valenciana (GVA) under project *Deep learning for adaptive and multimodal interaction in pattern recognition (DeepPattern)* (grant agreement PROMETEO/2019/121). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research; and the translators and project managers from Pangeanic for their help with the user study.

References

Akabe, K., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). Discriminative language models as a tool for machine translation error analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1124–1132.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González-Rubio, J., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Sanchis-Trilles, G., and Tsoukala, C. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Aziz, W., Castilho, S., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In *In proceedings of The International Conference on Language Resources and Evaluation*, pages 3982–3987.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.

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

Biçici, E. and Yuret, D. (2015). Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(2):339–350.

Bojar, O., Buck, C., Callison-Burch, C., Haddow, B., Koehn, P., Monz, C., Post, M., Saint-Amand, H., Soricut, R., and Specia, L., editors (2013). *Proceedings of the Eighth Workshop on Statistical Machine Translation*.

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

Castilho, S., Resende, N., Gaspari, F., Way, A., ODowd, T., Mazur, M., Herranz, M., Helle, A., Ramírez-Sánchez, G., Sánchez-Cartagena, V., Pinnis, M. a., and Šics, V. (2019). Large-scale machine translation evaluation of the iADAATPA project. In *Proceedings of the Machine Translation Summit*, pages 179–185.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367.

Daems, J. and Macken, L. (2019). Interactive adaptive smt versus interactive adaptive nmt: a user experience evaluation. *Machine Translation*, pages 1–18.

Dale, R. (2016). How to make money in the translation business. *Natural Language Engineering*, 22(2):321–325.

Denkowski, M., Dyer, C., and Lavie, A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404.

Domingo, M., García-Martínez, M., Estela Pastor, A., Bié, L., Helle, A., Peris, Á., Casacuberta, F., and Herranz Pérez, M. (2019a). Demonstration of a neural machine translation system with online learning for translators. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–74.

Domingo, M., García-Martínez, M., Peris, Á., Helle, A., Estela, A., Bié, L., Casacuberta, F., and Herranz, M. (2019b). Incremental adaptation of NMT for professional post-editors: A user study. In *Proceedings of the Machine Translation Summit*, pages 219–227.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., and Germann, U. (2014). The matecat tool. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132.

Freitag, M., Grangier, D., and Caswell, I. (2020). Bleu might be guilty but references are not innocent. *arXiv:2004.06063*.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471.

Green, S., Heer, J., and Manning, C. D. (2013a). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448.

Green, S., Wang, S., Cer, D., and Manning, C. D. (2013b). Fast and adaptive online training of feature-rich translation models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 311–321.

Guerberof, A. (2008). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus*, 7(1):11–21.

Hu, K. and Cadwell, P. (2016). A comparative study of post-editing guidelines. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 34206–353.

Jia, Y., Carl, M., and Wang, X. (2019). Post-editing neural machine translation versus phrase-based machine translation for english–chinese. *Machine Translation*, pages 1–21.

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

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the Association for the Computational Linguistics*, pages 67–72.

Koponen, M., Salmi, L., and Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, pages 1–30.

Kothur, S. S. R., Knowles, R., and Koehn, P. (2018). Document-level adaptation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73.

Lam, T. K., Kreutzer, J., and Riezler, S. (2018). A reinforcement learning approach to interactive-predictive neural machine translation. In *Proceedings of the European Association for Machine Translation conference*, pages 169–178.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Neubig, G., Dou, Z.-Y., Hu, J., Michel, P., Pruthi, D., and Wang, X. (2019). compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41.

Ortiz-Martínez, D. (2016). Online learning for statistical machine translation. *Computational Linguistics*, 42(1):121–161.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Peris, Á. and Casacuberta, F. (2019). Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language. In Press.*

Peris, Á., Cebrián, L., and Casacuberta, F. (2017). Online learning for neural machine translation post-editing. *arXiv:1706.03196*.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Pym, A., Grin, F., Sfreddo, C., and Chan, A. (2012). The status of the translation profession in the european union. Technical report.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Toral, A. (2019). Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

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

Turovsky, B. (2016). Ten years of Google Translate.

Vasconcellos, M. and León, M. (1985). SPANAM and ENGSPAN: machine translation at the pan american health organization. *Computational Linguistics*, 11(2-3).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Wuebker, J., Simianer, P., and DeNero, J. (2018). Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 881–886.

Appendix A User Questionnaire

How satisfied are you with the translation you have produced?

- Very satisfied.
- Somewhat satisfied.
- Neutral.
- Somewhat dissatisfied.
- Very dissatisfied.

Would you have preferred to work on your translation from scratch instead of post-editing machine translation?

- Yes.
- No.

Do you think that you will want to apply machine translation in your future translation tasks?

- Yes, at some point.
- No, never.
- I'm not sure yet.

Based on the post-editing task you have performed, how much do you rate machine translation outputs on the following attributes?

	Well below average	Below average	Average	Above average	Well above average
Grammatically					
Style					
Overall quality					

Based on the post-editing task you have performed, which of these statements will you go for?

- I had to post-edit ALL the outputs.
- I had to post-edit about 75 % of the outputs.
- I had to post-edit 2550 % outputs.
- I only had to post-edit VERY FEW outputs.

Based on the post-editing task you have performed, how often would you have preferred to translate from scratch rather than post-editing machine translation?

- Always.
- In most of the cases (75 % of the outputs or more).
- In almost half of the cases (approximately 50 %).
- Only in a very few cases (less than 25 %).

Which of the tasks do you think was the one that contained online learning? *(Note: This question was only asked after both tasks had been completed.)*

- Task 1.
- Task 2.

Give your opinion about the task you have performed.