# Machine translation of anonymized documents with human-in-the-loop

**Konstantinos Chatzitheodorou**
**M.ª Ángeles García Escrivà**
**Carmen Grau Lacal**
Pangeanic
Av. Corts Valencianes, 26
Bloque 5, 46015 Valencia
{k.chatzitheodorou, ma.garcia, c.grau}@pangeanic.com

## Abstract

In this paper, we introduce a workflow that utilizes human-in-the-loop for post-editing anonymized texts, with the aim of reconciling the competing needs of data privacy and data quality. By combining the strengths of machine translation and human post-editing, our methodology facilitates the efficient and effective translation of anonymized texts, while ensuring the confidentiality of sensitive information. Our experimental results validate that this approach is capable of providing all necessary information to the translators for producing high-quality translations effectively. Overall, our workflow offers a promising solution for organizations seeking to achieve both data privacy and data quality in their translation processes.

## 1 Introduction

Almost five years ago, the European Union, setting a milestone for data protection, enforced the General Data Protection Regulation (GDPR). Private and public organizations were required to remove sensitive content from public distribution involving European citizens under this legislation (European Parliament and Council of the European Union, 2016).

Text may need to be anonymized before it is translated to protect sensitive or confidential information. Text anonymization is a critical step in protecting sensitive or confidential information before machine translation (MT). Anonymization involves removing or disguising personally identifiable information or other sensitive data in a text to protect the privacy and confidentiality of individuals or organizations mentioned in the text (Pilán et al., 2022).

Anonymization is particularly important in situations where the translated text may be viewed by individuals who are not authorized to access the sensitive information contained in the original text. For example, in the case of medical records or legal documents, it may be necessary to remove personally identifiable information to protect patient or client privacy (Papadopoulou et al., 2022).

Moreover, post-editing machine-translated texts is often required to ensure that the translation accurately conveys the intended meaning and tone of the original text. A human-in-the-loop workflow for post-editing machine-translated texts can improve the quality of the final translation by leveraging the strengths of both human and MT (Lee et al., 2021).

By anonymizing the text before translation and utilizing post-editing workflows, the confidentiality and privacy of sensitive information can be maintained, while allowing the text to be effectively translated and used for its intended purpose. Furthermore, MT incorporates the factor of speed, meaning that post-editing is faster than translating from scratch.

In this paper, we propose a human-in-the-loop workflow for post-editing machine-translated documents that have been anonymized to protect sensitive information. The proposed workflow leverages the strengths of both humans and MT to improve the quality of the final translation while ensuring that the privacy and confidentiality of sensitive information are maintained.

## 2 Challenges of translating anonymized texts

Translating an anonymized text from one language to another can present some unique challenges for both an MT model and/or a professional translator. According to a study by Forsyth and Lam (2014), anonymized text may not provide enough context for the translator to accurately understand the meaning of certain words or phrases. This can lead to errors or inaccuracies in the translation. Anonymization can also result in the loss of information that would normally be useful for translation. For example, if a document contains references to specific cultural or historical events, these may be removed or obscured during the anonymization process. This is supported by research from Ruiz (2020).

Anonymized text may include non-standard language or jargon that is not commonly used in the target language. This can make it more difficult for the translator to find accurate translations for certain words or phrases. According to a study by Nemeskey (2020), non-standard language is one of the major challenges in MT. Some languages have more complex grammar and syntax structures than others, which can make it more difficult to translate anonymized text accurately, as pointed out by Renduchintala and Williams (2021).

In addition to language-specific challenges, translating texts may also require an understanding of cultural differences between the source and target languages. For example, if the original text includes references to cultural practices or beliefs that are not familiar to the translator, this can lead to inaccuracies in the translation. This is supported by research from Pratiwi (2022). Replacing the name of a location with a different one in order to achieve pseudo-anonymization could potentially cause cultural problems and misunderstandings, such as replacing "New York" with "Luxembourg". These two locations have very different cultural contexts and characteristics, hence the translator might lead to a more redundant target text.

Overall, translating anonymized text can be a complex and challenging process that requires careful attention to context, language, and culture. By understanding the unique challenges involved and using appropriate tools and techniques, translators can work to produce accurate and high-quality translations of anonymized text.

## 3 Related Work

An important area of research in MT is the development of techniques to handle sensitive or confidential information, such as medical records, legal texts, or bank documents. After conducting a thorough review of the relevant scientific literature in this field, it appears that no similar research has been carried out. Despite the absence of similar studies, researchers endeavor to enhance the power of MT to translate confidential information through the utilization of dictionaries and terminologies, as demonstrated in the works of Kirchhoff et al. (2011) and Zeng-Treitler et al. (2007). Nevertheless, none of these studies involve the inclusion of human intervention in the process. Conversely, there are some efforts from Computer-Assisted Translation (CAT) tools, such as XTM cloud,[1] that allow for the post-editing of anonymized texts. However, in the process, these tools replace named entities with numerical codes, which can sometimes cause confusion for translators and machines. An example of this type of anonymization can be seen in how the original text "John Smith is a professor at Stanford University" is transformed into "1 is a professor at 2". Such anonymization methods can pose a challenge for both human and MT models in comprehending the text. One alternative approach could involve substituting the original text with labels such as "NAME", "LOCATION", etc. Although this method may be superior to using codes, it still lacks vital details, such as whether the "NAME" label pertains to a male or a female.

Our research distinguishes itself from previous efforts by involving professional translators in the workflow to ensure that machine-translated output meets the standards of human translation. By working with meaningful sentence context and replacing sensitive information with fake data, both human translators and MT models can reduce the risk of errors and decrease the amount of time required for post-editing. This unique approach provides valuable insights into the role of human participation in MT and highlights the importance of considering human involvement in the development and implementation of AI-based technologies.

[1] https://xtm.cloud/

## 4 Workflow

In the context of our study, we introduce a workflow that combines the benefits of both humans and MT. It focuses on preserving the privacy and confidentiality of sensitive information while ensuring the accuracy of the final translation.

The proposed workflow involves several key steps, including the initial MT of the anonymized text, followed by a human post-editing stage. The post-editor reviews a pseudo-anonymized machine-generated translation and makes the necessary corrections to ensure that the translation is accurate and conveys the intended meaning. Then, the pseudo-anonymized text is replaced by the machine-translated text of the original text.

For instance, consider a case where a medical report needs to be translated for a patient who is traveling to a different country for treatment. The report contains sensitive medical information that needs to be anonymized before translation. In this scenario, the anonymization process may result in the replacing of personal names, medical facility names, and location information with labels (e.g., "NAME", "LOCATION", etc.) or with alternatives (e.g., "Angela" will be replaced with "Maria", "London" will be replaced with "New York", etc.). As a result, the MT may generate text that lacks contextual information, making it challenging for the reader to accurately understand the intended meaning. Following the anonymization of the text, a professional translator performs a post-editing task to ensure that the machine-generated translation accurately conveyed the intended meaning of the original text. The post-edited text is then subject to a final step, where an algorithm is used to replace the anonymized entities with their original versions in the translated text.

By employing this approach, sensitive information is protected, and patient privacy is maintained throughout the translation process. In addition, the use of pseudo-anonymization eliminates biases, while allowing for accurate and contextually appropriate translations.

Following is a high-level overview of the post-editing workflow for anonymized text:

- *Pseudo-anonymization*: The original text is processed to remove any sensitive information that may be present, such as names, addresses, or personal identifiers. To perform this task, we used Pangeanic's AI-driven Masker[2], which utilizes advanced techniques to automatically detect and replace sensitive personal data, such as names, addresses, or personal identifiers, within the original text. As part of our study, we leveraged the Faker library to pseudo-anonymize the sensitive information found in the documents. The Faker Python library (version 9.1.4) allows us to generate realistic and anonymized data by creating fake names, addresses, and other personally identifiable information (Faraglia and other contributors, 2014). We extended this, by utilizing the Genderize Python library (version 0.3.1), which uses probabilistic methods to predict the gender of a given name, enabling us to replace it with another name of the same gender (Ehrhardt and other contributors, 2018). By employing this technique, the context required for an MT to comprehend and accurately translate the text is retained to the greatest extent possible.

- *MT*: The anonymized text is fed into an MT system to generate a preliminary translation. This step provides a starting point for the human post-editor to work from. Our research methodology is designed to be flexible to meet the varying needs of our study. To achieve this, we support both in-house MT frameworks (e.g., ChatGPT-powered MT (OpenAI, 2022), OpenNMT (Klein et al., 2017), Marian (Junczys-Dowmunt et al., 2018), etc.) and publicly available providers such as Google Translate[3] to generate translations.

- *Human Post-Editing*: Professional translators or post-editors review the machine-generated translation and make the necessary corrections to ensure that the translation accurately conveys the intended meaning. They work to ensure that the translation is grammatically correct, contextually appropriate, and free of errors.

- *Entity Replacement*: In the final step, an algorithm is employed to replace the anonymized entities in the post-edited text with their original versions in the translated text. This

[2]https://pangeanic.com/data-masking-tool
[3]https://translate.google.com/

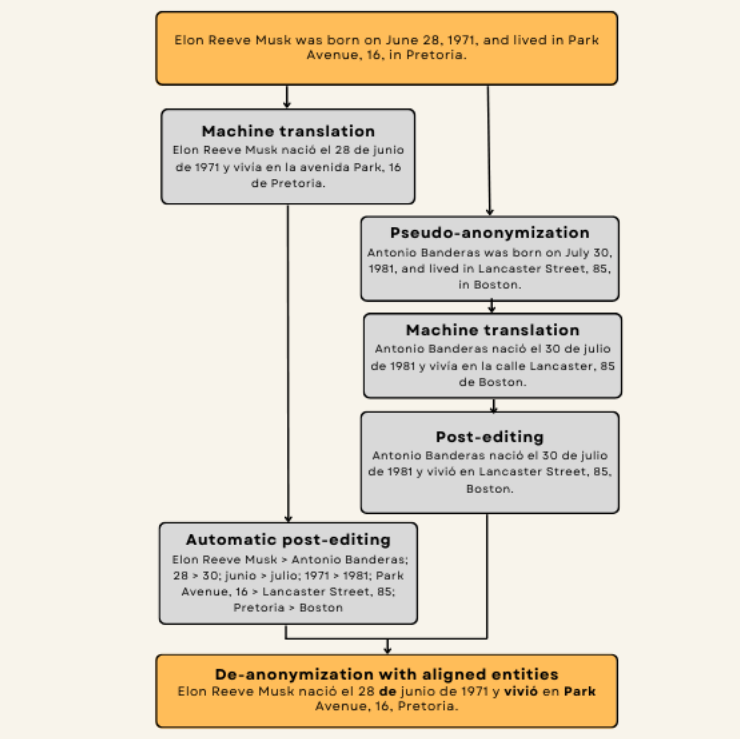


**Figure 1:** Workflow for MT of anonymized documents with human-in-the-loop

step ensures that the final translation is a faithful representation of the original text. To carry out the data replacement process, we use the Awesome aligner (version 2.2) to align the words/phrases between the original source and the machine-translated text (Dou and Neubig, 2021). This allows us to identify corresponding word pairs and accurately replace the pseudo-anonymized data with the machine-translated data of the original sentence.

Figure 1 illustrates the overall process flow of the architecture we have designed for MT of anonymized documents with human-in-the-loop. This architecture includes several components that work together to achieve this objective.

The workflow can be further customized based on the specific needs of the project and the type of sensitive information present in the original text. It allows for accurate and contextually appropriate translations while preserving the privacy and confidentiality of sensitive information. It can also be enhanced by integrating CAT tools with it.

Overall, the proposed workflow provides an effective solution for translating anonymized text while preserving the privacy and confidentiality of sensitive information. The use of both human and MT ensures high-quality translations that convey the intended meaning, which is particularly important in domains such as healthcare, legal, and financial sectors, where accuracy and confidentiality are critical.

## 5 Evaluation and results

To evaluate the effectiveness of the proposed human-in-the-loop workflow for post-editing anonymized texts, we conducted a series of experiments using both objective and subjective measures. The crucial steps of our workflow are (1) the pseudo-anonymization of the entities with fake entities and (2) their replacement with the machine-translated versions of the original entities after the post-editing process. The evaluation was conducted by assessing individual sentences.

For the subjective evaluation (step 1), we conducted a user study in which 14 participants of different nationalities (with a background in translation or linguistics) were asked to select which of the generated sentences better conveyed the original text. The participants had to choose among three options: the text that included pseudo-anonymized entities, the substitution with numeric codification, or the labeling codification. After carrying out the first part of this study, the participants were asked to provide their insights about the different methodologies used

to anonymize the original text and the issues identified during the task concerning the post-edition of the different alternatives.

The test set used in our study comprised a diverse range of documents, including legal contracts, medical reports, and financial statements. To ensure a representative sample, we sourced the documents from multiple industries and geographic regions, resulting in a test set that was both comprehensive and challenging. In total, it contained 100 sentences with an average length of 15 words per sentence. The shortest sentence in the test set was 3 words long, while the longest sentence had 45 words. Our test set consisted of various types of entities including 60 person names [PER], 80 locations [LOC], 20 organizations [ORG], 30 dates [DATE] in different formats, 20 bank account numbers [IBAN], 30 ID or passport numbers [ID], 60 telephone numbers [TEL] with or without country codes, and last 80 email addresses or URLs [EMAIL]/[URL], including subdomains, all of which were carefully annotated for an accurate analysis. We took steps to ensure that the test set did not contain any duplicated entities, to prevent any potential bias or skewing of results.

The results of this subjective evaluation show that the pseudo-anonymized text and the labeling codification were considered the most appropriate options even though some issues were highlighted. When analyzing the answers, we found out that several subjects chose multiple options, pseudo-anonymization, and labeling codification being the most frequent. After checking the comments, we realized that some of the issues could be avoided by using different post-processes after the pseudo-anonymization is performed.

A list of pros and cons for each of the main options selected is provided below. In addition, some of the above-mentioned issues and the corresponding post-processes suggested to avoid the problem are explained too. Probably, new processes will arise once the workflow is used in Production.

*Pros and cons of using pseudo-anonymization*

**Pros:**

- Sentences anonymized using fake entities instead of categories are more fluent and readable.
- Original entities replaced with fictional ones retain the meaning better.

**Cons:**

- Numeric ranges substitution could be unrealistic. For instance, *7 out of 5*. A possible post-process could be applied to force the second number of the range to be always higher than the first.
- It can be misleading if the fake entity has nothing to do with its context.

*Pros and cons of using labeling codification*

**Pros:**

- It provides a description of the replaced information without the actual details.
- It is possible to understand the original meaning.

**Cons:**

- Some labels are not clear enough. For instance, [DATE] may stand for a year only or a specific day of the month, etc. An option to improve the result could be replacing the format of the [DATE] label by providing different date formats, such as “MM, DD, YY”; “MM, YY”; “YY”; “DDMMYY”, or others.
- The lack of specificity may cause confusion.
- Different types of data are included in the same label. For example, the span “Director-General of the World Health Organization” was replaced by [JOB]; however, this span includes more than a job specification. Therefore, it would need to be split into two different tags [JOB]+[ORG]. For this type of issue, a new taxonomy matching a deeper level of detail would be necessary.

Regarding the objective evaluation (step 2), the participants were provided with different post-edited alternatives of an original text which included the machine-translated entities replacement. Each alternative results from a different anonymized option (anonymization with numeric codes, labeling codification, or pseudo-anonymization). They were first anonymized, machine-translated, and then, post-edited, and

| ORIGINAL SENTENCE | NUMERIC CODIFICATION | PSEUDO-ANONYMIZATION | LABELING CODIFICATION |
|---|---|---|---|
| To contact the Office of Scientific Integrity, call (404) 639-7570 or send an email to OADS@cdc.gov. | To contact the {1}, call {2} or send an email to {3}. | To contact the Office of Foreign affairs, call (345) 636-7545 or send an email to dfg@ghu.gov. | To contact the [ORG], call [TEL] or send an email to [EMAIL]. |

**Table 1:** Example of an evaluated sentence with different anonymization types.

finally, the entities were replaced with machine-translated ones.

Considering minor mistakes those which do not affect the meaning (grammar, word order, etc.), and major mistakes those affecting the meaning (mistranslation, omission, addition, etc.), the subjects had to rate the quality of each resulting translation based on the following scale:

- 2 or more fatal mistakes = 1 point
- 1 fatal mistake or >2 minor mistakes = 2 points
- 2 minor errors = 3 points
- 1 minor error = 4 points
- no errors = 5 points

The results shown in Table 2 indicate that the text with pseudo-anonymized entities received higher ratings compared to the text with numeric code or labeling substitutions. According to the participants' evaluations, the replacement of sensitive information with codes or labels did not preserve the meaning of the sentence completely and was rated lower in terms of quality.

The table indicates that the text with pseudo-anonymized entities received significantly higher ratings (mean = 4.33) compared to the text with other codifications (labeling mean = 4.14, and numeric codification mean = 3.91), which suggests that the pseudo-anonymized entities better preserved the meaning and characteristics of the original text.

The primary objective of these evaluations was to determine whether the pseudo-anonymized entities preserved the full meaning, i.e., gender and other characteristics of the original text. This evaluation enabled us to ensure that the pseudo-anonymized entities did not introduce any unintended biases or distortions to the original text.

As part of the second step of our evaluation process, we asked 5 professional translators to post-edit the pseudo-anonymized versions of the original text into Spanish and German. Following this, our algorithm replaced the pseudo-anonymized entities with the machine-translated versions of the original text. As mentioned above, the resultant output was verified by them, who examined whether the de-anonymized version was linguistically proficient as if they themselves had translated the anonymized entities. This process allowed us to validate the effectiveness of our methodology and assess its suitability for the study. By verifying the data replacement, we were able to identify any areas for improvement and refine our approach to ensure its accuracy. Results provided us with valuable feedback on the strengths and limitations of our methodology, enabling us to develop a more reliable and effective approach for future research in this area.

In general, the translators provided us with positive feedback for all the target languages. For Spanish, it was reported that pseudo-anonymization was clear enough to produce a correct and accurate text which always kept the intended meaning after replacing the anonymized text. The other two anonymization options introduced sometimes misleading information. For instance, in one of the sentences a nationality had been anonymized with the label [COUNTRY], which caused a concordance issue in the final version of the translation. For German, the reported observations were similar to those for Spanish. In this case, a problem related to pronoun use and inflection was reported due to the anonymization of "Thames". When using the label [LOC] or a numeric code, there was no information about the type of place, while with pseudo-anonymization, the post-editor got "Seine" instead, and could choose the proper pronoun and article, as well as their correct declination.

Overall, the experimental results demonstrate that the proposed human-in-the-loop workflow for post-editing anonymized documents can significantly improve translation quality while reducing the workload of human post-editors. Although

| Type of anonymization | Total points | Mean |
|---|---|---|
| Pseudo-anonymization | 303 | 4.33 |
| Labeling codification | 290 | 4.14 |
| Numeric codification | 274 | 3.91 |

**Table 2:** Ratings of Texts with Pseudo-anonymized Entities and Code Substitutions

our workflow has yielded promising results, it is important to acknowledge the risk that machine translation may not accurately capture the intended meaning of entities in the original text, which could result in mistranslations. Furthermore, the automated alignment process may also be prone to inaccuracies, which could further compound these risks.

## 6 Conclusion

In conclusion, the human-in-the-loop workflow for post-editing anonymized documents offers a promising solution for organizations seeking to balance the competing demands of data privacy and data quality. Our research represents a significant innovation in the field of MT and post-editing, as it utilizes cutting-edge techniques and is the first of its kind to be presented. By leveraging the strengths of both MT and human post-editing, our workflow enables efficient and effective translation of anonymized texts while preserving the confidentiality of sensitive information. Our experimental findings indicate that our approach is effective in reducing the risk of a human translator accessing sensitive information during the translation process.

We hope that our work will inspire further research on this topic and contribute to the development of more robust and efficient workflows for post-editing anonymized texts with human involvement.

## References


Dou, Zi-Yi and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Ehrhardt, Erica and other contributors, 2018. *Project description*. Retrieved February 24, 2023 from https://pypi.org/project/genderize/.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *OJ L 119, 4.5.2016, p. 1—88*.

Faraglia, Daniele and other contributors, 2014. *Faker's documentation*. Retrieved February 24, 2023 from https://faker.readthedocs.io/en/master/.

Forsyth, Richard S. and Phoenix W. Y. Lam. 2014. Found in translation: To what extent is authorial discriminability preserved by translators? *Literary and Linguistic Computing*, 29(2):199–217, 05.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344.

Kirchhoff, Katrin, Anne M Turner, Amittai Axelrod, and Francisco Saavedra. 2011. Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association*, 18(4):473–478, 04.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Lee, Dongjun, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. 2021. Intellicat: Intelligent machine translation post-editing with quality estimation and translation suggestion. *CoRR*, abs/2105.12172.

Nemeskey, Dávid Márk. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

Papadopoulou, Anthi, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. *arXiv*, arXiv:2205.06895v1.

Pilán, Ildikó, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Pratiwi, Putu Ayu Asty Senja. 2022. Translating and Interpreting in Intercultural Communication: A Study of Public Service Translators and Interpreters in Japan. *English Education: Journal of English Teaching and Research*, 7(2):157–168, Oct.

Renduchintala, Adithya and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender. *CoRR*, abs/2104.07838.

Ruiz, Nicolás. 2020. A general cipher for individual data anonymization. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 28(5):727–756.

Zeng-Treitler Q, Goryachev S, Kim H Keselman A Rosendale D. 2007. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007 Oct 11, pages 846–50.