NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Position paper: should machine translation be labelled as AI-generated content?

Simard, Michel

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp. 119-129, 2024-09

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=0c983431-cbf1-4a5c-bc9a-a3e3dc55c8cd
https://publications-cnrc.canada.ca/fra/voir/objet/?id=0c983431-cbf1-4a5c-bc9a-a3e3dc55c8cd

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

Position Paper: Should Machine Translation be Labelled as AI-Generated Content?

Michel Simard michel.simard@nrc-cnrc.gc.ca
National Research Council Canada

Abstract

In September 2023, the Government of Canada issued a *Guide on the Use of Generative AI* with recommendations for Canadian government institutions and their employees. As other similar documents published by various organizations in recent years, this document makes recommendations regarding transparency, stating that whenever generative AI is used to produce content, the reader should be informed that "*messages addressed to them are generated by AI*". While this guide does not address specifically the case of machine translation, it does mention translation as a potential application of generative AI. Therefore, one question that naturally arises is: Should machine-translated texts be explicitly labelled as AI-generated content wherever they are used? In this position paper, we examine this question in detail, with the goal of proposing clear guidelines specifically regarding MT, not only for government institutions, but for anyone using MT technology. Our main conclusion is that machine-translated text is indeed AI-generated content. As such, it should be explicitly marked everywhere it is used. We make recommendations as to what form this labelling might take. We also examine under what conditions MT labelling can be removed or omitted.

1 Introduction

In September 2023, the Government of Canada issued a *Guide on the Use of Generative AI*[1], providing "*preliminary guidance to federal institutions on their use of generative AI tools*". Among other things, this document makes recommendations regarding transparency, stating that whenever generative AI is used by Canadian government institutions, the users should be informed that "*messages addressed to them are generated by AI*" (Government of Canada, 2023).

While the recommendations in this guide are very general and do not target any one specific application of generative AI, the authors explicitly mention language translation as a potential use of these technologies. But they don't go as far as identifying machine-translated text as AI-generated content. Of course, machine translation and artificial intelligence are very tightly linked, both historically and technologically. Therefore, one question that naturally arises is: **Should machine-translated text be explicitly labelled as AI-generated content wherever it is used?**

In this position paper, we examine this question in detail, with the goal of proposing clear guidelines specifically regarding machine translation, not only for government institutions, but for anyone using MT technology to produce versions of a text in a language other than the one in which it was initially written. We present the wider context in which this question arises in Section 2, then address our fundamental question, as well as several others in Section 3. We wrap up with a summary of recommendations in Section 4.

Our main conclusion is that machine-translated text is indeed AI-generated content. As such, it should be explicitly marked everywhere it is used. We make recommendations as to what form this labelling might take. We also examine under what conditions MT labelling can be removed or omitted.

[1] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations...

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada
Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

2 The Wider Context: Labelling of AI-generated Content

In its recommendations to Canadian federal institutions on the use of generative AI technologies, the *Guide on the Use of Generate AI* includes guidelines on transparency: "*To maintain public trust and ensure the responsible use of generative AI tools, federal institutions should [...] identify content that has been produced using generative AI [and] notify users that they are interacting with an AI tool*". In particular, the guide contains a section about "*[d]istinguishing humans from machines*", with specific recommendations to "*[i]nform users when messages addressed to them are generated by AI*" or to "*use watermarks so that users can identify content generated by AI*" (Government of Canada, 2023).

These recommendations from the Government of Canada are not an isolated case. Increasingly, there has been pressure on Big Tech and media to label AI-generated content (abbreviated *AIGC* henceforth). The UNESCO's 2022 *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2022), which was adopted by all 193 member states in November 2021, includes a clause about identifying AIGC:

> 127. Member States should ensure that users can easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics [...]

This principle is gradually taking the form of laws and regulations in various places. For example, the European Commission (EC) added labelling requirements to its *Code of Practice on Online Disinformation* (European Commission, 2022), a voluntary framework of industry self-regulation to fight disinformation, to which most AI actors have already agreed (one notable exception is X, formerly Twitter, which pulled out during the summer of 2023). The EC's Digital Services Act (DSA) (European Commission, 2024) includes provisions requiring large online platforms to label "*manipulated audio and images*" (O'Carroll, 2023; Zakrzewski and Lima-Strong, 2023).

The United States, Canada and other countries are expected to adopt similar rules shortly. The United States government has recently convened the major players in AI to adhere to a set of guidelines aimed at ensuring safe, secure, and trustworthy AI. These guidelines specifically include provisions to "*[d]evelop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content.*"[2] A similar Code of Conduct (Innovation, Science and Economic Development Canada, 2023) was unveiled in Canada in September 2023, instructing companies to "*[d]evelop and implement reliable and freely available methods to detect content generated by their systems (e.g., watermarking)*" and to "*[e]nsure that systems that could be mistaken for humans are clearly and prominently identified as AI systems*" (Thompson, 2023; Pisano, 2023).

While legislation and recommendations have mostly focused on audio and video content, it is clear that textual content should not be an exception. Generative AI tools based on Large Language Models (LLMs) have rapidly been identified by analysts as a potential risk for the massive increase of dis- and misinformation. This topic was specifically addressed during a workshop organized by OpenAI, Georgetown University's Center for Security and Emerging Technology and the Stanford Internet Observatory in October 2021.[3] During this workshop, the question of Digital Provenance Standards was specifically discussed (Goldstein et al., 2023).

The primary goal of AIGC labelling is to fight dis- and misinformation, but there is a more general intention to inform users and create a more healthy and transparent social environment where trust can flourish, reflecting the ethical imperative "*to not blur the distinction between the categories of human and machine*" (Grinbaum and Adomaitis, 2022).

3 The Case of Machine Translation

In this document, we ask whether machine-translated texts should be labelled as AI-generated content wherever they are used. But before we can address this question, we first need to clarify whether MT *is* AIGC. Then, assuming it is, whether it is appropriate to label it, and if so, where and how. We address each of these questions (and more) be-

[2] https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-...

[3] https://openai.com/research/forecasting-misuse

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

low.

3.1 Is Machine Translated Text "AI-generated Content"?

TL;DR: Yes.

The Canadian Federal Government's *Guide on the Use of Generative AI*, as other similar documents, cites language translation as an example of tasks that generative AI can be used to perform or support. But it doesn't go as far as saying that machine-translated text is AI-generated content. We examine this question here.

When asked this precise question, it is surprising to see how many MT researchers and practitioners are hesitant or even reluctant to answer unequivocally. In practice, many of these people don't see themselves as "doing AI". They typically view their activity or area of expertise as either computational linguistics, natural language processing, machine learning, etc., but not AI. When pressed, many will highlight that AI is a badly defined concept to start with (more on this below), that it is a catch-all term or worst, just a buzzword. It is open for debate whether this tendency to dissociate with the "AI" label is the result of a conscious decision or just the self-preservation instinct of those who have survived a couple of AI winters!

Researchers have been studying the problem of automatic language translation for more than 60 years. But since the beginning, MT research was not only an end in itself: over time, it proved a fertile ground for the development and testing of some of the central ideas and methods of the artificial intelligence landscape: language analysis, understanding and generation, knowledge representation, pattern recognition, machine learning and, more recently, neural networks and deep learning, to name just a few. Today, the methods used for most text-generation AI applications are increasingly similar to those used for MT: the Transformer neural networks used in the vast majority of LLMs were first developed for translation (Vaswani et al., 2017). In some cases the tools (models, etc.) are literally the same: conversational AI systems are now increasingly used to translate text between many languages (Jiao et al., 2023; Hendy et al., 2023).

But does that make MT "artificial intelligence"? To answer this question, we also need to ask: What is Artificial Intelligence? In their classic AI textbook, Russell et al. (2010) cite no less than eight definitions of AI, among which the two following:

- "*The art of creating machines that perform functions that require intelligence when performed by people*." (Kurzweil et al., 1990)
- "*The study of how to make computers do things at which, at the moment, people are better*." (Rich and Knight, 1990)

Britannica, the web version of the well-known encyclopedia, describes AI as: "*the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings*."[4] And more recently, Coursera, the online learning platform, talks about "*the theory and development of computer systems capable of performing tasks that historically required human intelligence*".[5]

Admittedly, these are very broad and vague definitions. Yet it seems fairly uncontroversial to claim that translation is a task "*that historically required human intelligence*", that is "*commonly associated with intelligent beings*", at which "*at the moment, people are better*" or that "*require[s] intelligence when performed by people*."

To sum up: whether we look at it from a historical, technological or theoretical perspective, MT is AI, and therefore MT text is AI-generated content.

3.2 Should MT be labelled as AI-generated content?

TL;DR: Yes.

If MT text is AIGC, then it follows that any policy for AIGC should apply to MT as well. In the eyes of many, however, not all AI is made equal. For example, most would agree that a picture that was "enhanced" using a cell phone's AI-based photo improvement app doesn't quite fall into the same category as a photo-realistic image generated from a text prompt by a deep learning, text-to-image model.[6] In the case of machine translation, the text is generated from an input text which we assume was itself written by a human. Because the translation aims at

[4] https://www.britannica.com/technology/artificial-intelligence

[5] https://www.coursera.org/articles/what-is-artificial-intelligence

[6] Interestingly, we can ask whether photo-realistic image generation *is* AI, by any of the definitions in Section 3.1.

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

rendering the meaning of the source text in the target language as accurately as possible, it is tempting to see MT as merely an intermediary in the communication, a kind of "filter".

However, as argued earlier, the methods used for MT and those used for other text-generation applications are increasingly similar, when they are not altogether the same. Therefore, the risks inherent to the use of machine translation are essentially the same as those typically associated with chatbots and other conversational generative AI applications. There is a growing body of reported cases of MT errors with potentially grave consequences for people. Of course, there is the infamous example of the man who was arrested by the police in 2016, after Facebook's MT translated his "Good Morning" post to "hurt them" in English and "attack them" in Hebrew (Hern, 2017). But much more recently, Meta's MT was reported to add the word "terrorist" to some Palestinian users' Instagram profiles (Taylor, 2023). Again, in recent news, there were numerous reports of asylum applications being mishandled by United States immigration as a result of their over-reliance on MT (Liebling et al., 2020; Bhuiyan, 2023; Deck, 2023). MT systems routinely used by medical doctors in the United States when interacting with patients who don't speak English (Mehandru et al., 2022) have been shown to produce errors in medical documents, some of which can cause harm to patients (Khoong et al., 2019; Mehandru et al., 2023).

In summary, there is no good reason to believe that the nature of the risks inherent to MT are substantially different from that of those feared in other AIGC applications. Therefore, we recommend that MT should be labelled as AI-generated content.

3.3 Should there be a specific "MT" label?

TL;DR: Yes.

One of the purposes of labelling AIGC is, as UNESCO puts it, to "*ensure that users can easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics*". In the case of MT, however, there are additional reasons to do that.

Even though the quality of translations produced by MT has greatly improved over the past few years, systems are still known to produce errors. In general, MT quality is highly dependent on text domain and genre, but more importantly on the specific language pair involved (Hendy et al., 2023). Even for well-resourced language pairs such as English-German, accuracy errors (word or phrase mistranslations) make up the majority of problems and are still more than three times more likely to appear in state-of-the-art MT output than in translations produced by professional translators (Freitag et al., 2021). As pointed out by Vieira (2020): "*MT has great potential to facilitate and promote multilingualism, but its speed and usefulness may also prompt end-users to underestimate the complexities of translation while overestimating the capabilities of the technology, which in turn may lead to its misuse*." This has prompted the need for a new type of digital literacy, what has been called *MT literacy*: "*Using machine translation is easy; using it critically requires some thought. When faced with free, online machine translation, the important question is not how to but rather whether, when, and why to use this technology*." (Bowker, 2019b; Bowker and Ciro, 2019)

But MT literacy itself is useless if users are unaware that a particular piece of text is machine-translated. Therefore, for users to develop and make use of their judgement when dealing with MT output, it must be explicitly marked as such.

3.4 What should an "MT label" look like?

Obviously, an MT label should state that a given text is a translation and that the translation was performed by a machine. But any additional information that can help the user better assess the risks associated with MT can be useful. For example, an MT label might include:

- The source language. Knowing the source language may alert the user to specific kinds of errors, and therefore help them better assess the risk.
- The source text or, more conveniently, a pointer to the source version of the text. The user may be fluent enough in the source language that they prefer reading the original. Or they might know someone who does and who might be willing to verify the translation. Or they might have other tools at their disposal to help them assess the quality of the translation.
- The name of the MT system that generated it (possibly a detailed signature). Knowing what

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

system was used might help the user better assess the risks.

- A timestamp indicating when the translation was generated. Knowing that a translation was produced a (relatively) long time ago may inform the user about the expected quality.
- A full disclaimer or warning, informing the user of the potential risks associated with the technology, and/or pointing to resources on MT literacy.

Putting all this information inside a label might be cumbersome. A better approach would be to have a short text advising that "This text was translated from ⟨SourceLanguage⟩ by AI", followed by a link or reference to a document with more detailed information. The label should be written in the same language as the text itself. However, a link or reference to a version of the disclaimer in the document's source language could also be useful.

For organizations with large publication bases, it might be relevant to design a standard logo to accompany the label, thus making the texts (and their associated risks) easier and quicker to recognize by users.

3.5 Labelling vs. Watermarking

TL;DR: Labelling.

In the conversation about transparency and AIGC, there sometimes appears to be some confusion between *labelling* and *watermarking*. It is important to distinguish between the two.

According to Wikipedia: "*A label (as distinct from signage) is a piece of paper, plastic film, cloth, metal, or other material affixed to a container or product, on which is written or printed information or symbols about the product or item. Information printed directly on a container or article can also be considered labelling. [...] Labels may be used for any combination of identification, information, warning, instructions for use, environmental advice or advertising.*"[7] Disregarding the physical medium on which labels are said to be printed ("*...a piece of paper, plastic film, cloth...*") this definition fits nicely with what most people have in mind when talking about labels for AIGC.

Watermarking is quite a different beast. Initially, the term *watermark* referred to a recognizable image or pattern in paper used to determine authenticity. The concept was ported to the digital world in the form of "digital watermarks", i.e. markers covertly embedded in digital content. Digital watermarks have been used for a wide range of applications, such as copyright protection, source tracking, ID card security, fraud and tamper detection, etc.

While this sort of digital watermarking has been more commonly applied to audio, video or image data, techniques also exist for the watermarking of AI-generated text. For example, as early as 2011, Venugopal et al. (2011) proposed a watermarking method for statistical MT that operated by biasing the text generation towards a given portion of the lexicon, i.e. by favouring certain words over others. Text generated in this fashion could then be identified with high accuracy using a statistical test that "knew" the details of the bias. Similar methods have now been proposed for general neural text-generation applications (Kamaruddin et al., 2018; Kirchenbauer et al., 2023). Alternatively, some are advocating for "*AI to have its own alphabet*" (Croll, 2023): under such a scheme, MT systems would naturally produce text using a dedicated character encoding that would uniquely identify its synthetic origin.

For technology provider, the main purpose of watermarking is to be able to detect AI-generated text, especially content that was generated using their own technology, to avoid the model degradation that comes from training on synthetic data (Alemohammad et al., 2024; Shumailov et al., 2024). Because of the requirements of this application, the watermarking techniques developed for MT and other text-generation are typically designed to be resistant to later transformations to the text, such as revisions or post-editing, at least up to a certain point. As a result, a technique such as that of Venugopal et al. (2011) makes it possible to recognize MT text even if the text has been manually edited, for example by a translator. While this is an advantage for excluding MT data from future training sets and test data, it is problematic if the marking is required to be *reversible*, i.e. if we need to be able to "unmark" or "unlabel" text at will, as is the case here (see Section 3.8).

Another important requirement for the application we are interested in here is *perceptibility*: What-

[7] https://en.wikipedia.org/wiki/Label

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

ever form the labelling takes, the user must be able to see it (or hear it, sense it, etc.) somehow. While both types of watermarking above (encoding-based and lexical) can be detected using computer functions, they are not inherently perceptible.

A final problem with such watermarking is that it either disappears or becomes very difficult to detect as soon as the text is printed on a "hard" medium, such as a (paper) book, a restaurant menu, a road sign, etc. Similarly, watermarking may be lost to someone accessing the information through a screen reader, an audio recording, a braille reader or some other assistive technology.

In the end, for the purposes of informing the end user that a text was machine-translated, a textual label appears to be the simplest and most effective solution. This is what we recommend. However, depending on the intended use of the text, nothing precludes MT text to be both labelled *and* watermarked.

3.6 How do we know it's MT in the first place?

TL;DR: We don't, and so we must rely on voluntary identification.

3.6.1 Automatic Detection

Regulators (the EU, etc.) emphasize the responsibility of technology providers in developing ways to automatically detect AIGC. But most actors in the field recognize that building (and maintaining) such technology is a huge challenge, if not a losing battle altogether (Jovanović et al., 2024; Sadasivan et al., 2024; Krishna et al., 2023; Heikkilä, 2022). Some have recently proposed that any organisation developing a foundation model intended for public use (such as a LLM) should be required by law to demonstrate a reliable detection mechanism for the content generated by the model, as a condition of its public release, and make that detection mechanism freely available to users (Knott et al., 2023). While technology exists to do just that (see Section 3.5), no such legislation has appeared anywhere yet.

For MT, some people have looked at the problem of automatic detection in the past (see for example Bhardwaj et al. (2020)), and there are possibly some specific situations where it can be done reliably. For example, detection may be straightforward if watermarking has been used and the corresponding detection algorithm is available or when the challenge is to find out whether a specific MT system has been used, using methods similar to those developed for plagiarism detection (van der Werff et al., 2022). But for the general case, automatic detection of MT is probably not a viable approach.

3.6.2 Voluntary Labelling

Regarding general AIGC, early actions on the publishing side of Big Tech (social media, etc.) have focused on voluntary labelling by content producers (Suciu, 2023). For example, TikTok is encouraging users to label their AI-generated content as such (Sato, 2023), and Google and Meta require disclosure of AI content in political ads (Duffy, 2023; Isaac, 2023). At the AIGC-producing end, OpenAI puts restrictions on what can and cannot be done with their products. Their Usage Policies have requirements of transparency for some specific usages, encouraging users to "*disclose to people that they are interacting with AI*".[8]

Regarding MT, some MT providers (for example, Systran[9]) offer the possibility of including some form of labelling or watermarking in their system's output. But very often, MT is just one component within a larger application, and the MT system is not the one ultimately responsible for the display of its output.

Therefore, it should be the responsibility of whoever is disseminating (publishing, sending, posting, etc.) a machine translated text to propagate the label for that text if it already exists, or to create one if it doesn't.

It should be noted that this has implications for language service providers (LSPs): organizations who outsource the translation of their content to public or private-sector services will want to know whether their translation providers are using MT as a "productivity tool", and if so, whether all translated content has been manually verified and edited as required (see Section 3.8). Therefore, there should be a requirement for LSPs to appropriately label the texts they return to their clients.

But in the end, our recommendations apply to *all*: anyone disseminating machine translated texts or using MT to create content is responsible for labelling their content as appropriate.

[8] https://openai.com/policies/usage-policies/

[9] https://docs.systran.net/translate/en/user-guide/translation-tools/file-translation...

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

3.7 Where should MT be labelled?

TL;DR: Everywhere.

Should MT text be labelled everywhere it is used? Or should it be limited to institutional websites and other high-visibility communication channels with users? Should it apply only to contents with long shelf-life or should it also be used for punctual communication such as social media posts and institutional or commercial emails? What about institution-internal and personal communication: email, instant messaging, forms, software user interfaces, etc.?

There are clearly downsides to systematic, wall-to-wall labelling. Text is first and foremost a means of communication, and effective communication as is required from public and private institutions should be clear, precise and to the point. Labels may conflict with other visual requirements of the text, get in the way of communication and affect the user experience in unwanted ways. One extreme example that comes to mind is MT for software localization, where textual content often takes the form of individual words or phrases in buttons, menus, etc. Another example is column or row headers in tables or short captions in figures of automatically generated web pages. When these text items are machine-translated, it is not obvious how to label them clearly, especially if they are mixed with other, non-MT'd elements and if the labels should carry all the relevant information (see Section 3.4).

But then, how does one decide what to label and when? On the related topic of when and how MT text should be post-edited by professional translators, it has been suggested that the level of human intervention should relate to the purpose, value and shelf-life of the content (Way, 2013). Following this logic, labelling would be more appropriate for texts that are expected to have a longer shelf-life or are deemed to be more valuable or serve a more important purpose. But how do we measure value or purpose? And, perhaps more importantly, how do we measure the effect of translation errors on users? We have seen earlier how some errors can have serious consequences for users, even in short-lived, casual settings (see Section 3.2). It has also been observed that small errors, inconsequential in appearance, if they are recurring, may have just as serious effects on users as more critical errors, by gradually eroding the confidence of users over time. Research in the field of User Experience suggests that it may be useful to consider interactions with MT not only as static and isolated events but as part of a communicative process in the short and long term (Guerberof-Arenas and Moorkens, 2023).

Risks in translation (either human or computer) are a somewhat understudied area. In one of the few studies on the subject, Canfora and Ottmann (2018) hypothesize that in the realm of translations, as in areas where risks have been studied more systematically (healthcare, aviation, chemical industry, etc.), severe accidents are likely not caused by one single error but are the result of several failures, each of which would individually lead to only uncritical incidents. Furthermore, they suggest that all incidents, regardless of their severity, have the same root causes and that near misses are nothing but hazardous situations that only by chance did not turn into major accidents. One important implication of this observation is that we can effectively reduce the probability of severe accidents by reducing the number of near misses and minor incidents.

This suggests that the right way to go is to systematically label MT everywhere it is used. If the purpose of labelling is "*to not blur the distinction between the categories of human and machine*" (Grinbaum and Adomaitis, 2022), then this is the logical approach. For users who are knowledgeable about the limitations of MT, this will have the effect of "raising the right flags". And for those who are not, it will foster MT literacy by exposing users to "positive" and "negative" examples in various types of communication settings.

For situations where systematic and precise labelling might interfere with effective communication or with user experience, the best solution may be to have a general disclaimer at the top or bottom of the display (web page, document, form, etc.) stating that "some elements of this page may have been generated by MT", with a link or a reference to resources where the user may find more information. At the other end of the spectrum, for a very long machine-translated document, it may be necessary to repeat this disclaimer periodically.

3.8 What if the MT text was post-edited or verified by a human?

TL;DR: The label can be removed or omitted.

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

The question of transparency for AIGC is in large part one of responsibility and liability: If a person, either natural or juridical, is willing to stand behind a given content, i.e. certify or guarantee that this content is accurate, adequate, well-formed and fit-for-purpose, then in principle, this should suffice for that content to be viewed as "human-equivalent". And from there, remove any AIGC label that may have initially been affixed to that content, i.e. "unlabel" it.

In the case of MT, this "approval" amounts to certifying that the content is an accurate and well-formed ("fit-for-purpose") version in the target language of the source text of which it is a translation (Bowker, 2019a). In an ideal world, such verification would always be done by a professional translator, but in many practical settings, it can be performed by a competent, bilingual individual with a good knowledge of the original text domain, target audience and communicative intent of the translation.

"Responsibility" is the key word here: by removing an MT label or omitting to label MT-generated text, one is effectively taking personal responsibility for the validity and quality of that content (or, conversely, for any translation error that it may contain).

3.9 What if the source text that was machine-translated is itself AIGC?

TL;DR: It should be labelled as AIGC *and* MT.

This case is simple: any content naturally inherits characteristics of every step or processing it went through. So a text that was originally generated by an AI system (say, ChatGPT, as a response to a prompt or question) and then machine translated (either by the same system or a different one) remains fundamentally AI-generated. In principle, this kind of text would carry both labels: AIGC and MT. And to remove both labels, one would have to validate for each separately, i.e. make sure that the source text is factually accurate, grammatically correct, etc. and then make sure that its translation is adequate, grammatically correct, etc.

4 Summary of Recommendations

Machine translation is AI-generated content. As such, it is subject to the same recommendations as other AIGC. Our recommendation is that machine-translated text be systematically labelled everywhere it is used. The label should explicitly say that the content was machine-translated and specify the language from which it was translated; if possible, the label should also provide a link or reference to the original text, as well as pointers to general resources about MT literacy.

The only condition under which such a label could be omitted or removed is if the entirety of the affected text has been verified and certified fit-for-purpose, ideally by a professional translator, otherwise by a competent bilingual who accepts full responsibility for the quality and appropriateness of the translation.

Acknowledgements

This work has greatly benefited from conversations with and comments from many people. I wish to express special thanks to (in alphabetical order) Gabriel Bernier-Colborne, Lynne Bowker, Atsushi Fujita, Cyril Goutte, Rebecca Knowles, Samuel Larkin, Chi-kiu Lo, Alan Melby and Joss Moorkens.

References

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. (2024). Self-Consuming Generative Models Go MAD. In *The Twelfth International Conference on Learning Representations*.

Bhardwaj, S., Alfonso Hermelo, D., Langlais, P., Bernier-Colborne, G., Goutte, C., and Simard, M. (2020). Human or Neural Translation? In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6553–6564, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bhuiyan, J. (2023). Lost in AI translation: growing reliance on language apps jeopardizes some asylum applications. *The Guardian*. Thu 7 Sep 2023.

Bowker, L. (2019a). Fit-for-purpose translation. In *The Routledge handbook of translation and technology*, pages 453–468. Routledge.

Bowker, L. (2019b). Machine translation literacy as a social responsibility. In Adda, G., Choukri, K., Kasinskaite-Buddeberg, I., Mariani, J., Mazo, H., and Sakriani, S., editors, *Proceedings of the 1st international conference on Language Technologies for All*

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

(LT4All), pages 104–107, Paris, France. European Language Resources Association.

Bowker, L. and Ciro, J. B. (2019). *Machine translation and global research: towards improved machine translation literacy in the scholarly community*. Emerald Publishing, Bingley, first edition edition. OCLC: on1075580986.

Canfora, C. and Ottmann, A. (2018). Of ostriches, pyramids, and Swiss cheese: Risks in safety-critical translations. *Translation Spaces*, 7(2):167–201.

Croll, A. (2023). To Watermark AI, It Needs Its Own Alphabet. *Wired*. July 27, 2023.

Deck, A. (2023). AI translation is jeopardizing Afghan asylum claims. *Rest of World*. 19 April 2023.

Duffy, C. (2023). Google to require disclosures of AI content in political ads. *CNN*. Fri September 8, 2023.

European Commission (2022). The 2022 Code of Practice on Disinformation | Shaping Europe's digital future.

European Commission (2024). The Digital Services Act package | Shaping Europe's digital future.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv:2301.04246 [cs].

Government of Canada (2023). Guide on the use of generative AI.

Grinbaum, A. and Adomaitis, L. (2022). The ethical need for watermarks in machine-generated language. arXiv:2209.03118 [cs].

Guerberof-Arenas, A. and Moorkens, J. (2023). Ethics and Machine Translation: The End User Perspective. In Moniz, H. and Parra Escartín, C., editors, *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer International Publishing, Cham.

Heikkilä, M. (2022). How to spot AI-generated text. *MIT Technology Review*. 19 December 2022.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210 [cs].

Hern, A. (2017). Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*. Tue 24 October 2017.

Innovation, Science and Economic Development Canada (2023). Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems.

Isaac, M. (2023). Meta to Require Political Advertisers to Disclose Use of A.I. *The New York Times*. 8 November 2023.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., Shi, S., and Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745 [cs].

Jovanović, N., Staab, R., and Vechev, M. (2024). Watermark Stealing in Large Language Models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22570–22593. PMLR.

Kamaruddin, N. S., Kamsin, A., Por, L. Y., and Rahman, H. (2018). A Review of Text Watermarking: Theory, Methods, and Applications. *IEEE Access*, 6:8011–8028.

Khoong, E. C., Steinbrook, E., Brown, C., and Fernandez, A. (2019). Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*, 179(4):580.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A Watermark for Large Language Models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Eyers, D., Trotman, A., Teal, P. D., Biecek, P., Russell, S., and Bengio, Y. (2023). Generative AI models should include detection mechanisms as a condition for public release. *Ethics and Information Technology*, 25(4):55.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 27469–27500. Curran Associates, Inc.

Kurzweil, R., Richter, R., Kurzweil, R., and Schneider, M. L. (1990). *The age of intelligent machines*, volume 580. MIT press Cambridge.

Liebling, D. J., Lahav, M., Evans, A., Donsbach, A., Holbrook, J., Smus, B., and Boran, L. (2020). Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Mehandru, N., Agrawal, S., Xiao, Y., Gao, G., Khoong, E., Carpuat, M., and Salehi, N. (2023). Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Mehandru, N., Robertson, S., and Salehi, N. (2022). Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2016–2025.

O'Carroll, L. (2023). Google and Facebook urged by EU to label AI-generated content. *The Guardian*. 5 June 2023.

Pisano, V. (2023). How can we tell whether content is made by AI or a human? Label it. *Macleans.ca*. 29 May 2023.

Rich, E. and Knight, K. (1990). *Artificial intelligence Subsequent Edition [M]*. McGraw-Hill College, second edition.

Russell, S. J., Norvig, P., and Davis, E. (2010). *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, 3rd edition.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2024). Can AI-Generated Text be Reliably Detected? arXiv:2303.11156 [cs].

Sato, M. (2023). TikTok introduces a way to label AI-generated content. *The Verge*. 19 September 2023.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Suciu, P. (2023). 'Created By AI' Warning Labels Are Coming To Social Media. *Forbes*. 2 August 2023.

Taylor, J. (2023). Instagram apologises for adding 'terrorist' to some Palestinian user profiles. *The Guardian*. 20 October 2023.

Thompson, E. (2023). Ottawa unveils new AI code of conduct for Canadian companies. *CBC News*. 27 September 2023.

UNESCO (2022). Recommendation on the Ethics of Artificial Intelligence.

van der Werff, T., van Noord, R., and Toral, A. (2022). Automatic Discrimination of Human and Neural Machine Translation: A Study with Multiple Pre-Trained Models and Longer Context. In Moniz, H., Macken, L., Rufener, A., Barrault, L., Costa-jussà, M. R., Declercq, C., Koponen, M., Kemp, E., Pilos, S., Forcada, M. L., Scarton, C., Van den Bogaert, J., Daems, J., Tezcan, A., Vanroy, B., and Fonteyne, M., editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. (2011). Watermarking the Outputs of Structured Prediction with an application in Statistical Machine Translation. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Vieira, L. N. (2020). Machine translation in the news: A framing analysis of the written press. *Translation Spaces*, 9(1):98–122.

Way, A. (2013). Emerging use-cases for machine translation. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Zakrzewski, C. and Lima-Strong, C. (2023). Europe moves ahead on AI regulation, challenging tech giants' power. *Washington Post*. 14 June 2023.

© 2024 His Majesty the King in Right of Canada, as represented by the National Research Council Canada