Position Paper

84000's Position on AI and the Machine Translation of Canonical Literature

June 2023

From its inception, 84000 has embraced technology as a major vehicle to realize its mission. 84000 conducted in-depth surveys on how best to represent texts in digital form, researched, defined and documented its own customization of text encoding initiative (TEI) schema so that the digital form could be consistently applied, created a robust platform for dynamic electronic publishing, and created an entire ecosystem of digital tools for translators. In the field of Buddhist digital humanities, 84000 is recognized as a significant contributor.

Naturally, 84000 is very interested in the opportunities presented by Artificial Intelligence (AI). The research and testing of machine learning, natural language processing, and specifically Large Language Models (LLMs), is developing at unprecedented speed. Of particular interest is Machine Translation (MT)—the translation of a text from one language into another language by a computer without any human involvement during processing—the evolution of which has made significant strides in recent times. Nowadays, MT is probability-based and relies on large corpora of source texts and their human translations as training data. It works best with pairs of modern languages for which enormous numbers of matching text passages, or at least vast quantities of digital text in both languages, can be used to train it. Until recently, the volume of available training data in Tibetan and Sanskrit was thought to fall far short of what was necessary for good results. However, recent advances have greatly enhanced what MT can do with less data, and the data-sets that comprise 84000's published translations have provided a significant and valuable contribution to researchers developing MT models for Tibetan language that yield readable translated text. These are not of the standards of accuracy reached for much more extensive corpora; thanks to advances in the language generation of good English, they read plausibly, but that very plausibility may disguise their lack of accuracy.

Machine translation, with its sophisticated processing of language based on calculating the statistical likelihood of a certain rendering, in ideal circumstances can give surprisingly readable results. Indeed, some proponents of machine translation tend to assume that MT has essentially solved the "problem" of translation. In reality, however, a complete match between one word in one language and its counterpart in another language rarely exists. And when it comes to different registers of language and the purpose of a text, things naturally get even more complex.

In any decision about whether to adopt MT or how to deploy it, it is first important to understand the expectations of different groups of users. A casual, private user who wants to understand the gist of something in another language does not expect 100% accuracy. In contrast, users such as companies or organizations responsible for providing translations for public distribution have different expectations (for example, ensuring that the output is reliable, appropriate, and nuanced). Users in this latter group have mostly concluded, through trial and error, that machine translation can only be reliably used in combination with human effort.

Concerns in the 84000 Context

Generally speaking, 84000 believes that the use of AI as a tool to assist translation has considerable value. Currently, though—and especially with languages as sparsely spread as Tibetan—the output of automated MT engines comes nowhere near matching the reliability and quality of competent work done by a qualified translator. Despite the sophistication of their calculations, AI systems work without "understanding" either the source text or their own output. Linguistically, they will be unable to cope properly with unusual terminology, syntax, and language that has not yet been captured by a set of tokens and calculations—which is not uncommon in the Tibetan canonical literature. And beyond the linguistic realm, a proper understanding of context, register, and purpose can only be brought to bear on the task of translation by experienced, qualified translators deeply familiar with the source material and the target language and culture.

84000's mission is to translate the Buddha's words from the Tibetan Buddhist Canon into modern languages, and to make them available to everyone. The texts we work with are considered incredibly precious and meaningful for many, and they have been through a venerable and careful process of oral transmission and then translation, in India, and then in Tibet, all before the 13th century. The tremendous value these Tibetan scriptures preserve relies on the care and precision of those transmissions. The teachings they contain are complex and profound, and their meaning is encoded in many subtle ways that require careful unpacking and judicious rendering in the English language. For a project that strives to translate this collection for the general, educated reader, ensuring that the quality and readability of the translations is maintained is non-negotiable. Furthermore, in translating the texts from Tibetan into English, 84000 relies heavily on the living tradition of oral explanation to understand their meaning: It is the still transmitted living tradition that brings the texts to life. This is why we believe that the transmission and communication of the Buddha's words in English would inevitably be impaired and confused by over-reliance on machine translation and particularly by the proliferation of unedited, unsupervised machine translations.

Concerns around the accuracy of MT aside, we must also ask ourselves to what degree we want humans to be involved in the things that matter deeply to us. For 84000, ethics and genuine intention remain of vital importance. In undertaking the translation of this collection, 84000 feels a deep responsibility not only to a vast, global community of scholars, practitioners, and interested readers in this generation and the next, but also to the living tradition itself, to those who uphold its lineage and who ensure the transmission of Dharma over generations. There is no certainty that ethical decision making can be reliably integrated into machine translation, and it is even more questionable whether LLMs will remain accountable to anyone or aligned with the goals and purposes for which they were created.

In our view, these deep questions and concerns make the autonomous machine translation of sacred texts ethically problematic. Simply put, 84000 believes that as an agent in the transmission of living wisdom, AI is neither reliable nor appropriate. We believe that the primary actor in the translation of sacred literature must always be the human mind.

The Opportunities for AI at 84000

84000's position today is that for reliable translation and the respect of dharmic intention, accountability, and ethics, these sacred texts should not be processed using machine translation alone. However, as long as trained and qualified human translators are involved as the primary agents, we favor the deployment of AI and other technology tools in the type of process often referred to as machine-assisted translation (MAT).

As 84000 continues to carefully select and fund numerous scholars and translators, we will also continue to develop machine-assisted translation tools and integrate them into the translation process. These tools, among other things, allow us to:

- Check translations for omissions and errors;
- Offer alternative translation choices to translators;
- Draw on other translators' work without laborious manual searches;
- Find related passages across the Tibetan, Sanskrit and Chinese source texts;
- Identify inconsistencies across translations, and
- Help locate people, places, and terms in the Tibetan corpus and map those to the translation corpus.

We are very excited about the increasing potential impact of MAT on the efficiency and quality of the human translation process. As we produce and publish a carefully curated collection of translations in multiple data formats, along with extensive footnotes, glossaries, cross-references, digital resources, and supporting documentation, 84000 is committed to a responsible and transparent process of translation: we remain respectful to the living tradition, and accountable for the translations we produce.

Looking forward, it is on the basis of this balanced view that our strategic planning will continue to explore how we can leverage technology to build the appropriate tools to support translators who are dedicated to the authentic translation and transmission of the Dharma.