GeekSpeak

Jost Zetzsche

Data and the Fine Print, or How to Create a Sh*tstorm

I like Twitter. It's a good way to learn what's happening and at the same time have an additional motivation to process and curate information so that you can share noteworthy articles and information yourself. It was in that spirit that I shared an article by Matthew Blake about the dangers of lawyers using Google Translate, specifically regarding quality and confidentiality.[1] Not only that, but I even tagged on a "Good read" to my tweet.

It's true that I had not noticed that the article was "sponsored content." (But truth be told, I have ghostwritten a number of articles for sponsored content placement and they were still pretty good, if I do say so myself.) Either way, I was not quite prepared for the storm that broke loose, a very small portion of which you can follow at the link provided at the end of this article.[2] I did not jump with both feet into the assumed controversy right away, but a few days after the original eruption, I revisited the contentious topic—the issue of confidentiality when using services like Google Translate and Microsoft Bing Translator—and was surprised by what I found.

Like probably most of you, I had always assumed that everything passing through one of those two services would be used by Google and Microsoft. Well, that is only partially true, but it's especially important for us to know the exact point when the data is being used.

When you go to the terms of service on Google Translate, it's exactly like Blake's article claims. Here is the language:

> When you upload, submit, store, send, or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations, or other changes we make so that your content works better with our Services), communicate, publish, perform publicly, display publicly, and distribute such content.[3]

This is pretty clear-cut and very much along the lines of what we expect. Your source content will be used (not your target, unless you use one of the tools on the site to modify the suggested translation).

Google Translator Toolkit, the minimalistic translation environment tool that Google offers, also uses your content, only here it uses both source and target:

> We may use the content you upload to Google Translator Toolkit to improve Google services pursuant to our Terms of Service [see above]. If you delete your content from Google Translator Toolkit, we will delete the content from our servers and, from that point forward, will not use it for any additional improvements to Google services.[4]

I'm not a lawyer, but in my mind the last addition means that while the data is not being processed anymore once you delete it, whatever has been gained from the data while you had it stored with Google Translator Toolkit will still be used.

However, things are different once you use Google Translate API. (We use API in most translation environment tools essentially any time we enter the "API Key" and have to pay for the use.) In that case:

> Google does not use the content you translate to train and improve our machine translation engine. In order to improve the quality of machine translation, Google needs parallel text—the content along with the human translation of that content.[5]

Now, I'm not sure about the parallel text statement. Statistical machine translation engines typically do use monolingual text alongside parallel text. I also do not know why they would need the monolingual content of the non-API Google Translate but not this one. But, hey, what do I know, right?

All this said, Google is assuring us that it will not use any of our data

Information and Contacts

The GeekSpeak column has two goals: to inform the community about technological advances and at the same time encourage the use and appreciation of technology among translation professionals. Jost is the co-author of *Found in Translation: How Language Shapes Our Lives and Transforms the World,* a perfect source for replenishing your arsenal of information on how human translation and machine translation each play important parts in the broader world of translation. Contact: jzetzsche@internationalwriters.com.

if we pay for the translation service. Did you know that? I didn't.

Let's move on to Microsoft and the data that gets submitted to Microsoft Bing Translator. Microsoft has all of its terms nicely put together on one page:

> Microsoft Translator does not use the text you submit for translation for any purpose other than to provide and improve Translator, including improvements to the quality and accuracy of translations provided by Translator. (…) The text we use to improve Translator is limited to a sample of not more than 10% of randomly selected, non-consecutive sentences from the text you submit, and we mask or delete numeric strings of characters and e-mail addresses that may be present in the samples of text. The portions of text that we do not use to improve Translator are deleted within 48 hours after they are no longer required to provide your translation. If Translator is embedded within another service or product, we may group together all text samples that come from that service or product, but we do not store them with any identifiers associated with specific users.[6]

Okay, kind of what we thought. What about Microsoft Translator Hub, the customizable machine translation engine that Microsoft offers?

> The Hub retains and uses submitted documents in full in order to provide your personalized translation system and to improve the Translator service. After you remove a document from your Hub account we may continue to use it for improving the Translator service.[7]

That is a little "less generous" than Google. Even after you withdraw your documents, they might still continue to be processed.

What's really interesting is that there is also an exception. Just like with Google, if you pay (enough) you can opt out of your data being processed. If you subscribe to a monthly volume of 250 million characters or more, you may request to have logging turned off for the text you submit to Microsoft Translator.

So, if you pay a little more than $2,000 per month ($2,055 to be exact[8]), you can *request* not to have your data processed by Microsoft to improve the translation service. (The same terms apply to Microsoft Translator Hub.)

So, to summarize, if you do not pay for either Google's or Microsoft's services, your data will be processed. If you pay (in Microsoft's case: if you pay a whole lot), your data will be left alone. That is at least what the legal language says. And that should have an impact on the ongoing discussions on confidentiality concerns when using generic machine translation services.

And Blake's article? He was essentially right, since he was not talking about professional linguists who would likely be using API Google Translate, but about the casual user in the legal field. His concerns about quality are spot on as well.

What about us not being in a position to have an impact on those matters? After I published an early version of this article in my newsletter, I sent it to one of the people at Microsoft who is responsible for the Microsoft Translator program. His response: "Looks like it is time to revise our behavior one more time." So, we can make a difference.

To put this all into perspective, here is an interesting outlook from a localization manager of a reasonably large information technology company with whom I have worked in the past. He wrote to me recently to share his concern about the decreasing quality of translation this past year, wondering aloud whether generic machine translation engines like the ones discussed in this article are to blame. When I shared this on Twitter, a deluge of responses suggested he should find new vendors or that it's the responsibility of the individual translator to choose a tool. I agree. Still, we would be wise to "treasure all these things and turn them over in [our] mind." ■

Notes

1. Blake, Matthew. "Man vs. Machine: Google Translate Jeopardizes Client Confidentiality," *Above the Law* (January 5, 2015), http://bit.ly/1EDfDmJ.

2. http://bit.ly/Jeromobot-twitter.

3. Google Terms of Service, www.google.com/intl/en/policies/terms.

4. Google's Policy on Sharing and Deleting Your Translation Data, http://bit.ly/1u2nV3V.

5. Google Translate API, Data Confidentiality, http://bit.ly/18KBT12.

6. Microsoft Translator Privacy Statement, http://bit.ly/Microsoft-privacy.

7. Ibid., http://bit.ly/Microsoft-privacy.

8. Microsoft Translator Data, http://bit.ly/1zPSRa7.