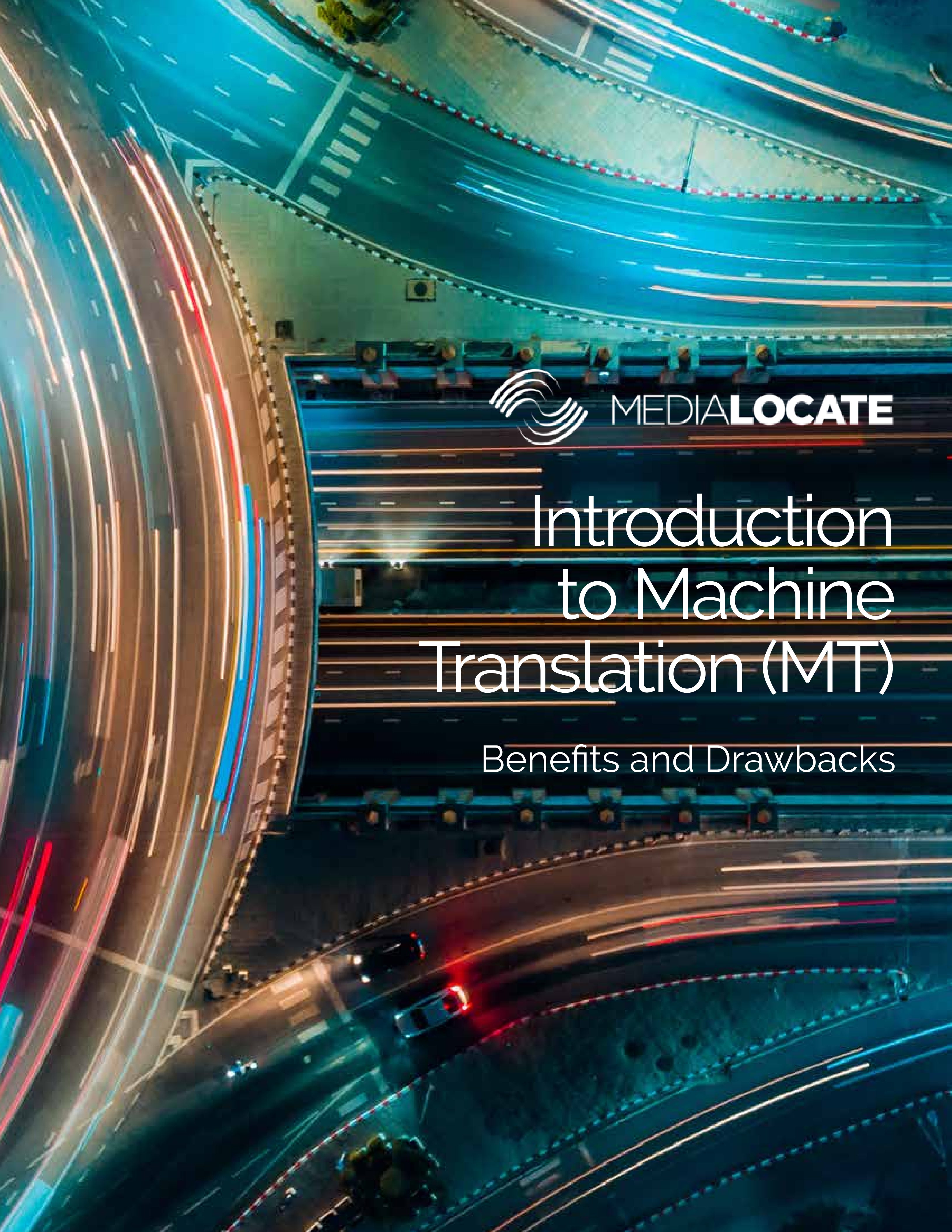


MEDIALOCATE

Introduction to Machine Translation (MT)

Benefits and Drawbacks

Introduction

This guide is meant as a short, basic introduction to Machine Translation -- its benefits, and its drawbacks. It is not meant as a comprehensive guide to all of the details and processes that need to be undertaken for a successful MT program. However, it contains valuable information for those who are considering MT as an option for localizing their content.

What is Machine Translation (MT)?

Machine translation (MT) is an automated translation process that can be used when a fully human translation process is insufficient in terms of budget or speed. This is achieved by having computer programs break down a source text and automatically translate it into another language. Although the raw computer translated content will not have the quality of materials translated by qualified linguists, the automated translations can be post-edited by a linguist in order to produce a final, human-like quality translation.

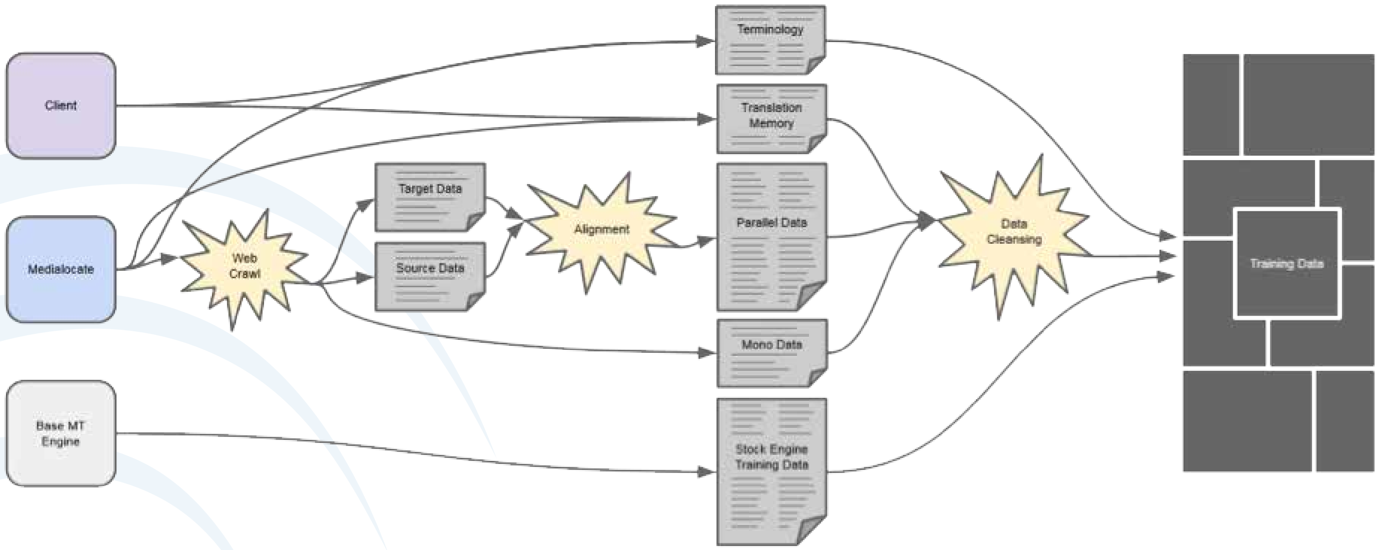
When combined with human post-editing (called Post Editing of Machine Translation or PEMT for short), MT can potentially reduce translation costs and turnaround times while maintaining a level of quality appropriate for the content's end-user. The increase in speed and the reduction of costs depends largely on the collaboration between linguists and engineers in order to train the computer system to translate for a specific language and domain. The more data the engineers have, such as translation databases translated by human linguists, normal language documents written in the source and target languages, or glossaries, the better the MT output will become. This, of course, leads to less time the post-editor will need to spend on editing the MT output and further reduces the time and costs for those translations.

It is important to understand how this training process works so that the MT engine that is built can give the best ROI and meet the needs of a translation program. Although MT can take some time and cost to be set up, the savings for translation programs that have a consistently large amount of content to translate can be very substantial. So what goes into this set-up and how can you make the most of an MT program?

What Are the Different Types of Machine Translation (MT)?

There are several types of MT systems, and the process for setting up a program varies a little with each. At MediaLocate we use an engine that has initial training data that is then refined by our engineers in collaboration with linguists in order to produce a high quality MT engine. The process works by taking data from as many sources as possible and feeding it into the MT engine so that it is able to learn the semantics of a language and the words that it should use while translating automatically.

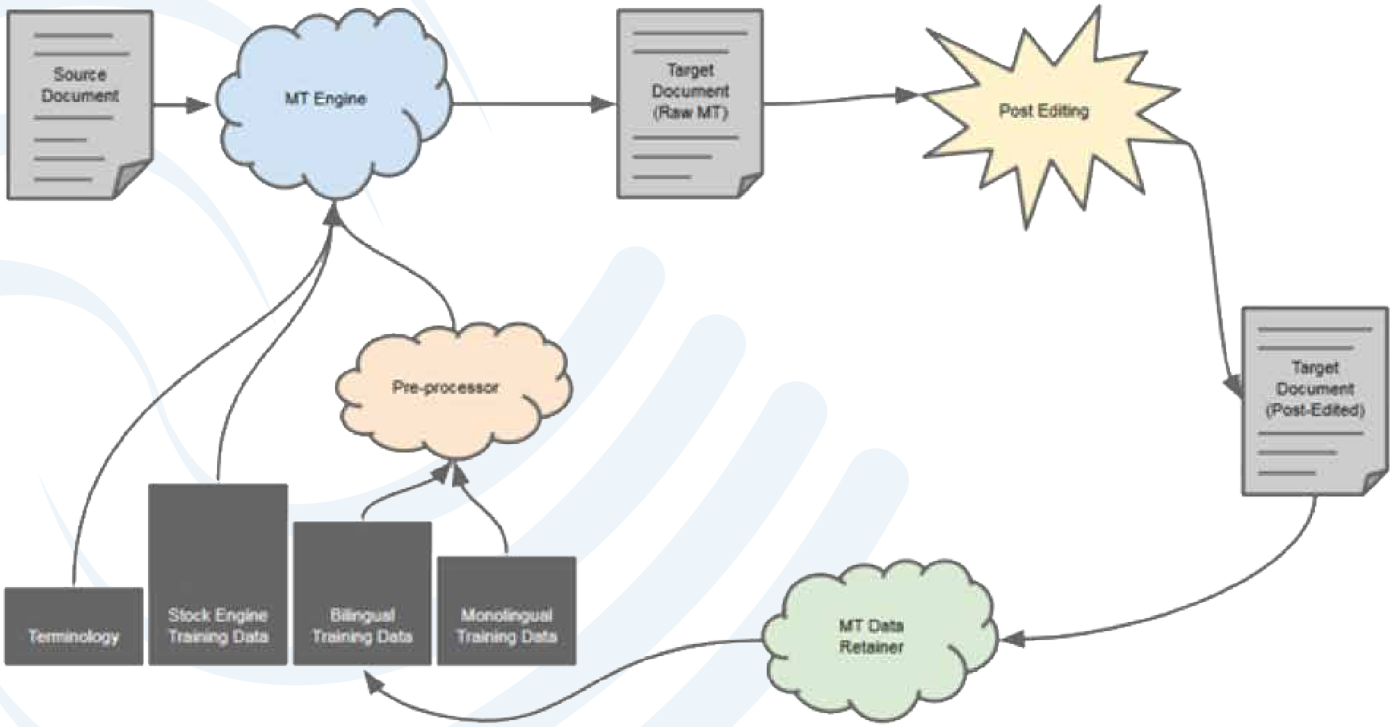
If we take a look at the following figure, we can see what kinds of data go into the initial training of the MT engine:



As you can see, a lot goes into the initial set up of a well-performing MT engine. The process requires the client requesting the translation to provide as much reference material as possible, e.g., specific terminology, existing glossaries, or translation memories containing past translations. MediaLocate also provides the translations we have previously completed for a client. Depending on the industry that the MT engine is being built for, we will seek out additional data on the web that identifies target and source data that we are able to align in to a source and target (translation) database creating parallel data. We also gather monolingual data, texts in the source and target language that do not have translations, which can help the MT engine understand and form sentences in those languages. The MT engine already has a baseline to start from or stock data for different industries or fields. The data is then put through data cleansing to remove any odd items such as corrupted text or items that will not help the MT engine or might confuse it in some instances.

The more data that is gathered and the greater the upfront effort put in here, the greater the reduction of time and cost can be for using PEMT. However, programs can normally start off with more limited data as long as they understand that the training of the engine will happen in real time. In other words, an MT engine might be built with limited data and initially, there might not be much of a reduction in time or cost with a low quality engine. As the post-editing linguist works, they are able to feed more and more data back into the engine, which will allow it to gradually increase its quality and efficiency, resulting in a gradual improvement in the turnaround times and costs.

Below, you can see how this process works where a source document is taken and translated by the engine and then post-edited by a trained linguist. After the content is post-edited, the MT engine then retains this data and adds it to the existing bilingual training data:



So while it is important to gather as much data as possible for the initial training of the engine, it is also possible to start with limited data and train the engine gradually with the understanding that, starting out, there will not be the same kind of improvements in costs and timelines as with a fully-trained engine. While all of this works well, it is worth inoting that there are some languages, industries, and types of translations that machines are not very good at!

What types of content are best for MT?

Before starting an MT program, it is important to consider the type of content to be translated. There are some types of documents that MT is not very good at handling and others which are quite suitable for MT. The key here is to ask, "How ambiguous is the text that is being translated?" By this, we mean "How many different ways could a text be understood or interpreted and how idiomatic is the text?" For example, if we saw the English phrase, "I spilled the beans there.", we would know that someone was conveying the message that they had let a secret or some important information slip out to someone whom it was not intended for. When a machine sees this phrase, it will translate it quite literally. However, in other languages the idiom 'spilled the beans' would obviously not have the same meaning and would thus be understood literally (i.e. someone knocked over a bowl with beans in it). This means that marketing texts, slogans, poetic texts, and literary prose are normally not good candidates for machine translation, as it is difficult for the machine to pick up on such intricacies and provide a correct and corresponding phrase in the target language.

In contrast, documents that use rather plain or structured language are ideal candidates for MT. Things such as technical publications, user guides, training materials, datasheets, legal documents, user-generated content, and quickly perishable, low visibility content in general are all ideal for MT, since they are generally not idiomatic and are straightforward or very structured, which makes it easy for the machine to identify patterns and provide a translation of acceptable quality. When there is a large amount of text to translate, it can often be advantageous to consider MT as a great way to save on time and effort. Many major corporations and international organizations have been using MT (often in combination with PEMT) to achieve speed and cost efficiency with these types of text, while maintaining a separate program to have human linguists translate their marketing materials.

Conclusion

From all of this we can see that there are some clear benefits, as well as drawbacks to MT. The training and set-up of an engine requires highly skilled engineers and linguists as well as some cost and time. Making this investment, though can have a huge benefit and ROI for all future translation projects and help to reduce the time needed to get new documents translated. The initial investement is actually minimal compared to the savings an MT program can yield. Also, while there are some types of content that even a trained engine will not be able to handle very well, when considering doing translations for large amounts of unambiguous text, MT can provide relatively good quality translations and help to keep costs down. So while it is not a silver bullet for any and all translation needs, it is good to consider MT for at least part of an overall translation strategy.

There are, of course many more intricacies and aspects of MT that have not been discussed here, but if you are interested in more information please contact our Machine Translation Department here at MediaLocate at **ml_mt_team@medialocate.com** and one of our MediaLocate team members will be happy to discuss the various aspects of MT with you.

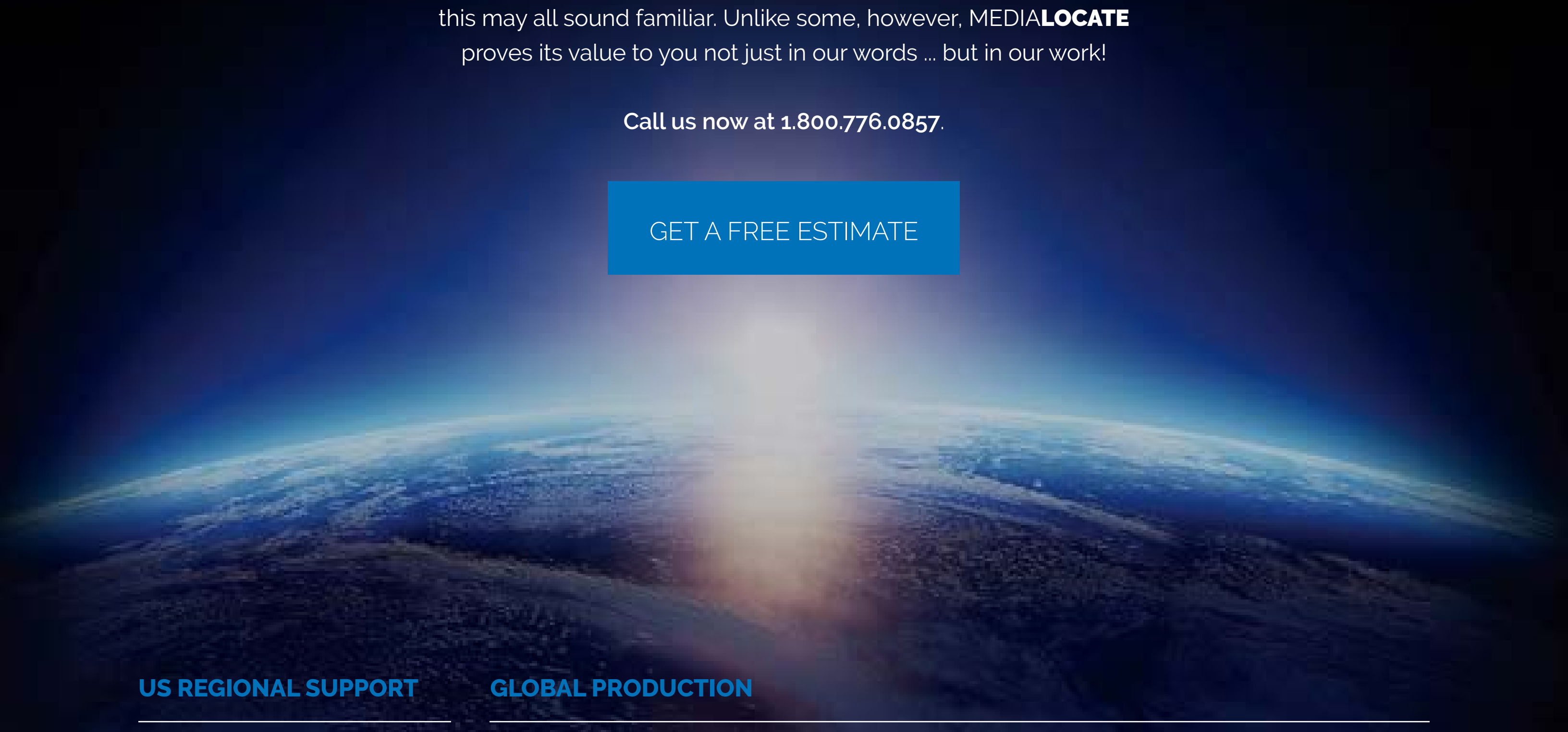
MEDIALOCATE

WE SPEAK HUMAN

If you've been searching awhile for the right language service provider, this may all sound familiar. Unlike some, however, MEDIA**LOCATE** proves its value to you not just in our words ... but in our work!

Call us now at 1.800.776.0857.

GET A FREE ESTIMATE



US REGIONAL SUPPORT

WEST COAST
995 Rosecrans Street
San Diego, CA 92106
Phone: 619-487-1394
SD_info@medialocate.com

EAST COAST
404 5th Avenue
New York, NY 10018
Phone: 609-216-5975
NYC_info@medialocate.com

GLOBAL PRODUCTION

SLOVAKIA
Podzamska 21
949 01 Nitra
Slovak Republic
slovakia@medialocate.com

RUSSIA
8 Rudnichniy Street
Stary Oskol
Belgorod
Russia, 309517
russia@medialocate.com

CHINA
Plaza 66 Tower II
1366 Nanjing Road West, 15th Floor
Shanghai, China 200040
shanghai@medialocate.com

SINGAPORE
#44-01 Suntec Tower One
7 Temasek Boulevard
Singapore 038987
singapore@medialocate.com

THAILAND
3703 B.B. Building (7th Floor)
54 Sukhumvit 21 Rd.
Bangkok 10110, Thailand
thailand@medialocate.com

KOREA
3rd Floor, Wonkwang Building,
283-4 Neung-dong, Kwangjin-gu,
Seoul, Korea
korea@medialocate.com

JAPAN
31F Osaka Kokusai Building
2-3-13 Azuchi-machi, Chuo-ku,
Osaka 541-0052, Japan
japan@medialocate.com