July 15, 2013

Do-it-yourself MT: Taking (statistical) machine translation to the next level

Uwe Muegge

Do-it-yourself MT: Taking (statistical) machine translation to the next level

New web-based statistical machine translation services are currently revolutionizing the market. These SaaS solutions allow users to customize an MT engine with their own translation memories. As most of these services follow the subscription model, launching a DIY MT project costs only a fraction of deploying a traditional machine translation tool, which makes this powerful technology affordable for even the smallest organization.

Image: © mrPliskin/ istockphoto.com

By Uwe Muegge

During the last two years, several SMT services have emerged that give users a high degree of control over the quality of machine-generated translations. Their business model is simple: Make a powerful statistical machine translation system accessible via the Web and allow users to customize an MT engine by uploading their own translation memories. And reduce the price of using a customized SMT system to a fraction of what high-quality machine translation traditionally costs.

Traditional approaches to customizing SMT

Google Translator Toolkit

Google Translate now supports more than 70 languages and has more than 200 million daily users, which makes Google's statistical machine translation the most widely used automatic translation service. In 2009, Google launched the Translator Toolkit, a free cloud-based translation memory tool that lets users post-edit translations generated by Google Translate. In fact, Google Translate lets users upload their own terminology databases, however, these glossaries can only be used for post-editing, not for customizing the machine translations Google Translate creates.

The only way users of Google Translator Toolkit can customize the output Google Translate produces is by re-using their previously post-edited SMT output. Google Translator Toolkit currently does not offer any functionality for customizing Google Translate beyond the ability to select a source and target language.

License SMT software

Up until recently, licensing a commercial statistical machine translation product was the primary solution for corporate users who wanted to take full advantage of SMT. Language Weaver, which was acquired by SDL in 2010, practically owned this market. As the name of the product indicates, the 'Enterprise Translation Server' was designed for large commercial and governmental entities who would engage Language Weaver to build custom SMT systems. In other words: The provider of the machine translation system would typically not only sell a translation software product but also customize the SMT system with either client- or domain-specific data.

The major drawback of this type of SMT solution was its high cost: Not only for licensing the software, but also for the high-end server hardware required to run this software, and the professional services involved in the customization process.

Moses

Moses started in 2005 as an academic experiment at the University of Edinburgh, and since then has evolved into a massive open source project with many institutional supporters throughout the world. In fact, as of 2012, the European Union funds the Moses Core project, which aims at making Moses' open source statistical machine translation accessible to an even wider audience. And this is where the challenge lies: While Moses is both free and supports any language combination, this statistical machine translation system is not exactly an easy-to-use application. Truth be told, Moses is not even an application but a collection of software modules designed to be run on a Linux server.

In other words: Moses is a powerful and very flexible SMT system that is supported by a broad and well-funded development community, but it takes a background in computational linguistics to operate Moses. So even though the Moses software is free, it's the staff and hardware infrastructure that's necessary to operate this MT system that makes using Moses a prohibitively expensive proposition for most small and medium-sized organizations.

The DIY approach to statistical machine translation

Using a cloud-based web service

When several companies started hosting their own SMT system on Web-enabled servers and made it easy for users to customize those MT engines, that was a major paradigm shift. With a hosted solution, users now no longer have to make a heavy up-front investment in either software , hardware, or human resources, but just pay for services as they go. In this type of Software as a Service (SaaS) model, the user does not have to deal with, or even understand, the complexities of statistical machine translation.

All users do in a DIY MT environment is upload their own data for training the SMT engine, upload the source documents for translation, and download the translated target documents. The only software the user interacts with directly is the SMT service provider's browser-based user interface, which is typically limited to simple project management functions.

Using your own data

In contrast to Google Translator Toolkit, which lets users generate and post-edit generic translations, DIY systems make it easy for users to customize the SMT engine. The purpose of customization is to have the statistical machine translation system generate translations that are consistent with the terminological

THREE DIY SMT SERVICES THAT MAKE IT EASY TO GET STARTED

Microsoft Translator Hub

Like the Translator Toolkit, Google's initiative to improve the quality of the Google Translator, the goal of Microsoft's Translator Hub is to take the Microsoft/Bing Translator to the next level. Unlike the Google Translator Toolkit, however, the Microsoft Translator Hub is a DIY SMT service that lets users create their own customized MT engines. Best of all: Using the Microsoft Translator Hub is completely free for users translating less than two million characters per month.

hub.microsofttranslator.com

KantanMT

KantanMT is a brand new DIY SMT service that just opened its doors. Unlike many other SMT service providers, KantanMT neither has up-front customization charges nor per-word translation fees. Instead, KantanMT charges a flat monthly subscription fee for customizing and maintaining either 20 or 40 engines for €499 and €999, respectively. KantanMT offers a free trial period of 14 days.

www.kantanmt.com

SmartMate

SmartMate is more than just a DIY SMT service; it offers a complete cloud-based translation management solution, including online editing and terminology management functionalities. Monthly subscription rates start at US$85 for 100,000 words and two editor seats. SmartMate offers a free trial period of five days.

www.smartmate.co

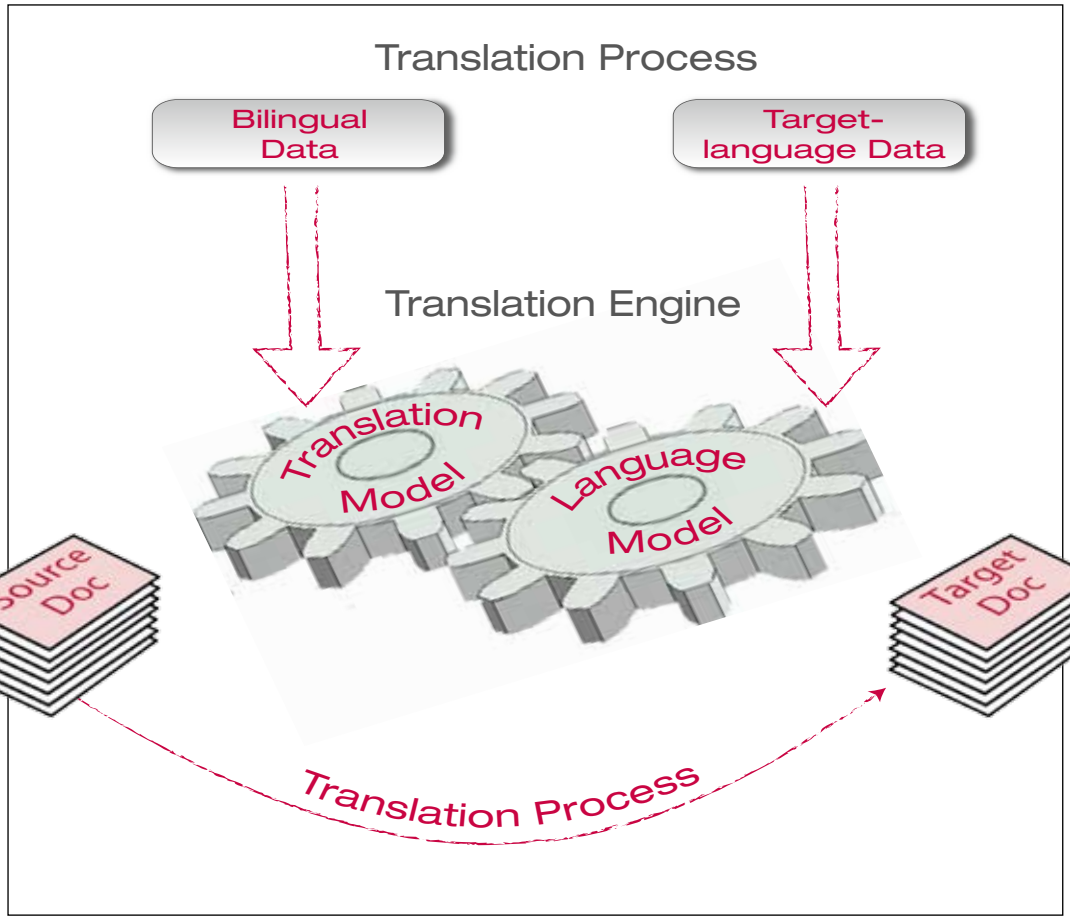

Figure 1: Functional diagram of how statistical machine translation works

and style conventions of their own organization. In other words: After customization, the translations the SMT system creates should be more usable and require less post-editing than translations created by traditional MT systems. Customization is typically a multi-stage process during which the user submits multiple sets of user-specific data. For training the translation model, which generates a set of possible translations, the user uploads a corpus of aligned sentences in the source and target language, typically in the form of a translation memory. For training the language model, which helps the SMT system pick the most 'fluent' translation generated by the translation model, the user uploads a corpus of translated sentences.

There are two things to know about SMT customization: The first is that customization requires fairly large data sets, with 10,000 sentences typically being the lower limit. This means that in order to create an SMT engine that produces translations that approximate your organization's style and terminology, the user must have a substantial set of available translations. And, in addition, these translations should have been reviewed for accuracy to ensure that the training data is of the highest quality. The other thing to know is the fact that SMT customization can be very granular. In other words: An organization can have multiple SMT engines, each for a specific domain. For instance, a user could have one engine for translating user manuals and another engine for translating.

Deploying a machine translation solution in days if not hours

As is true for many other Software as a Service (SaaS) solutions, setting-up a do-it-yourself SMT service typically takes much less time than rolling-out a traditional statistical machine translation solution. After creating an account, the main task is to upload the bilingual and monolingual sets of training data, and then wait for the customization process to complete.

Depending on the service provider and the size of the training corpora, customization can take anywhere from less than an hour to little more than a day.

In addition, getting a DIY statistical machine translation project started does not require the user to make any major financial or long-term contractual commitments. In fact, many providers of DIY SMT solutions let users try their services for free!

How to successfully roll-out your own DIY MT project

Manage your linguistic assets

You probably already know that translation memories are valuable assets: TMs help translation buyers avoid having to pay full price for the translation of material that was fully or even just partially translated before. In the context of statistical machine translation, translation memories become even more valuable as these bilingual documents, together with customer-specific translation-only documents, form the basis for customizing an SMT engine.

Currently, many – if not most – providers of DIY SMT services aim their offerings at small and medium-size language service providers. However, this type of solution is just as attractive to small and medium-size buyers of language services, provided these buyers have access not only to their translated documents, but also their translation memories.

Invest in training staff

As mentioned above, it doesn't take a computational linguist to operate a cloud-based do-it-yourself statistical machine translation system.

In fact, users don't even have to be translation experts to perform the basic functions of customization and translation in this type of environment.

However, organizations that wish to maximize the benefits of DIY SMT should consider hiring or training a translation or localization manager.

The translation manager will determine which projects go through a) a traditional human translation process, b) a pure SMT process, or c) an SMT process with subsequent human post-editing. In addition, it would be helpful to provide training to the organization's authors to enable them to use the same terms and the same style consistently. While consistent style and terminology in the source language help improve the quality of any translation project, stylistically and terminologically consistent source texts are particularly important in statistical machine translation projects.

Manage user expectations

Do-it-yourself statistical machine translation is a great new technology that makes high-quality automated translation accessible to a much wider audience than ever before. But if you think that DIY MT makes human translation obsolete overnight, think again. Statistical machine translation is a very powerful technology, but even the most advanced translation engine cannot make up for deficiencies in the source. If writers use sentence structures or terms that didn't occur in the training material an SMT system is based on, translation errors are inevitable.

On the other hand, DIY SMT is suitable whenever the alternative to machine translation is no translation at all, e.g. the translation of knowledge-base articles or user forum messages. And of course translations generated by a DIY SMT system make an excellent source for post-editing by human linguists.

Summary

Do-it-yourself statistical machine translation is a technology that holds a lot of promise - especially for organizations that currently use (S)MT and wish to improve translation quality. Cloud-based DIY SMT services not only make it easier to generate customized, or user-specific, machine translations; these services also make high-quality MT more affordable than ever before. In fact, several DIY SMT providers offer their services at such low cost that it puts this powerful technology within reach of even the smallest organization – including individual freelance translation/post-editing specialists.

contact

Uwe Muegge has more than 15 years of experience in the translation and localization industry, having worked in leadership functions on both the vendor and buyer side. Uwe has been with CSOFT International, a provider of language services based in Beijing, since 2008, and he currently serves as Senior Translation Tools Strategist for North America.

uwe.muegge@csoftintl.com
www.csoftintl.com