Google Translate: Everyone's Language Wallah

Introduction

Old joke: If you can speak three languages, you are trilingual. If you speak two languages, you are bilingual. If you speak one language, you are an American.

The problem is that other languages exist. In Brazil, educated professionals speak Portuguese and two or three other languages. For everyday work and communication, Portuguese is still the dominant language. The same fact holds true in most economic powerhouses. In order to keep pace with information available on Web sites, in Web logs, and even Tweets, translating a source language into my native language (English) is becoming more important. Even though I lived in Brazil and had a working knowledge of Portuguese, I need software safety nets.

Like many knowledge workers, I have relied upon software that takes a source language such as French, German, or Japanese and translates it into English. I have a number of translation resources. I use some open source tools built on the GNU gettext framework (http://www.gnu.org/software/gettext/); for example, code from Google's gettext commons (http://code.google.com/p/gettext-commons/). I have experimented with a range of shareware products. If you want to give some of these systems a trial run, navigate to ITShareware.com (http://www.itshareware.com/catlist-code_58-start_0-sort_0.htm).

When AltaVista.com was the big dog in search, I found the Babel Fish online translation service useful. Now part of the Yahoo suite of online services, Babel Fish (based on Systran's technology) can handle some light-weight translation tasks. If you are not familiar with this service, navigate to http://babelfish.yahoo.com/. I had to click past an add for Hot Fudge Brownies to access the service, but it is free and works reasonably well.

Today most organizations are like the fabled city of Babylon. In ancient times, so the legend goes, the residents of Babylon had a tough time communicating. Organizations have the same problem and not just with customers who speak a different language. Even small firms have to deal with an ever-growing flow of information which may be in Arabic or Ukranian. Most organizations struggle with a confusion of tongues, converting an office complex in San Jose to a mini Tower of Babel.

The need to process Web log content, email messages, and even 140-character Tweets increases the need for a swift, reliable automated way to translate electronic information. The brute force method once relied on human translators. Today, the flow of information makes the traditional methods too expensive and slow.

Machine translation systems (sometimes inaccurately labeled online translation systems or automatic translation systems) have been available for many years. These systems work

reasonably well, typically delivering a translation that can be used to get the gist of a source document. The machine translation systems often do a much better job with scientific, medical, and technical source documents than with more colloquial source material. General business writing falls somewhere in the middle in accuracy and usability.

Government entities have been important customers for vendors of machine translation technology. Intelligence agencies and research groups have an insatiable appetite for translations of a wide range of source content. The volume of content that requires translation continues to sky rocket. There is neither money nor sufficient human translators to keep up with the amount of material that must be translated. Machine translation now carries this burden.

The capabilities of machine translation systems often extend beyond taking a source document in French and generating a translation in English. Today's systems in use in certain government agencies perform translation and then identify themes, entities such as individuals and organizations, and concepts. Some systems perform additional text analysis to tag each component of a source document with a theme. A handful of vendors have ventured into sophisticated analytics that identify the sentiment or emotion expressed in a document and giving each document or "fact" in a document a reliability or confidence grade.

These advanced functions are an exciting area of research at universities and language research centers worldwide. The problem, of course, is that language continues to change. The nature of human communication manifests man's creativity. A group discussion becomes an online discussion and then morphs into a "WebEx" or "webinar". Software has to be quite intelligent to take "Facebook" in a company report and translate that into *réseau social*. The problem is that language is a moving target. Software can change but it often has some difficulty with idioms, neologisms, and words that seem to say one thing but mean quite another given the sender's and recipient's context. The problem is difficult, and many developers aim to provide systems that are "good enough" when installed. Licensees can add custom dictionaries and interact with the software system to "teach" it to handle certain types of content.

The Machine Translation Archive (http://www.mt-archive.info/) provides a useful resource to learn about the effort put into cracking the problem of cracking the information age's Tower of Babel problem.

I have licensed translation tools from Systran (http://www.systransoft.com), which bills itself as "the leading supplier of language translation software." I used Systran as my translation *wallah* (an Anglo Indian word for a person in charge of a task). The company offers a range of translation products for commercial and government use. A home office version of the translation system is available in language pair packs; for example, English to German and German to English. Systran supports 52 languages. If you want to experiment with the system, you can purchase a Web translator for about $50 or purchase a home-office suite for $800. The small business and enterprise versions are priced based on the language pairs you want to translate and the components such as work flow components you require to solve your organization's language problems.

My experience with Systran has been largely positive. I have a working knowledge of several languages. I can figure out an unclear passage by looking at the source document and the Systran version. When I need verification, I have to turn to a native speaker of the source language. Over the years, my need to tap a native speaker has decreased. In my opinion, the quality of machine translation systems has improved incrementally over the last five years. Most outputs are “good enough” for the type of work I perform.

When using machine translation systems to covert a document in English to a version in German, for example, close attention to word choice and sentence structure pays dividends. I get better results when I submit text written in short, declarative sentences. We have learned to avoid idioms, potential confusing homonyms like "wait" and "weight", and neologisms like the awful coinage “webinar”. Touching up is necessary.

Commercial machine translation software offers a wide range of options. For example, for me to add a double byte language to my basic Systran system, I need to purchase additional language packs. A “language pack” is essentially the knowledge base that makes it possible for the Systran software to convert a language. For example, the English World Pack which supports Chinese, Dutch, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, and Swedish, costs about $1,000. The company offers a discount for purchases of more than five licenses.

Systran, however, is a company that has had to deal with a somewhat unexpected disruption in the world of machine translation. In the last fiscal year for Systran, the company's revenues suffered a decline. In the most recent financial reports available for the French company, the firm continues to suffer in the present economic climate.

I am reluctant to make a cause and effect connection between the deterioration of Systran's revenues and the increased capabilities of Google's translation services, but Google's expanding machine translation capabilities may be a contributing factor.

Google’s translation services and its translation system are free as I write this in September 2009. Try the Google system by pointing your browser to http://translate.google.com. The first thing to note about the page is that unlike Yahoo’s Babel Fish, Google places no advertising between you and the translation system. Second, the user can provide a url, paste text into the translation box, or upload a document. There are 52 languages supported. In the tests I ran today, I did not encounter any document length limits.

The system supports multi-lingual search. You can enter a phrase in your native language. The Google system will translate your query into the target languages supported by the system. The search results will be returned regardless of the language of the page indexed by Google. You can translate pages in different languages using the Google Translate function. I saw a demonstration of a search system developed by Pertimm (http://www.pertimm.com/en/), located near Paris, that supported multi-lingual queries. Google’s implementation is a major advance in search functionality in my opinion.

There is a Translate option on the Google Docs menu bar as well. When I tested the service, I logged in and created a new document. I typed sample text into the document window. I then clicked on the Tools menu option and selected Translate. The Translated document appears in

the document window. The translate window offers two new options. First, you can replace the original document with this translation and copy the translation to a new document. One slick feature for me was that the translation preserves the formatting of the original document.

Google also offers a Translator Toolkit. You can find this by clicking on the link on the Google Translate splash page. These free components make it possible for a technically-savvy user to build a translation application using Google's system. The programming approach is consistent with the approach taken on Google's Code Playground. A person with elementary programming skills can explore the functionality of the Translate system. Google also provides a Gadget (code snippet for a Web page) so your Web pages can appear in other languages.

For more experienced developers, Google provides an AJAX Language API. This programming interface allows a developer to create an application that weaves together a number of Google services. For example, it is possible to build a complex application that translates and detects the language blocks of text within a Web page. In addition, a developer can enable translations on any text field or text area in a Web page. An example would be transliterating text to Hindi. The API makes it possible for users to spell out phonetically Hindi words using English and have them appear in the Hindi script. Google provides useful examples of its Translation API at http://code.google.com/apis/ajaxlanguage/documentation/#Examples You or your developers can give the Google Docs translation function a spin. You will need to get a free Google account.

The enhanced translation features in general and the Google Docs's Translate feature in particular are significant for three reasons. First, the service is available within a user-friendly word processor. This means that anyone with basic knowledge of word processing can create and translate documents with zero training. Google's interface is spare, but it makes a complicated task using traditional translation software almost foolproof.

Second, a developer can fiddle with Google's sample code and create specific applications or browser add ins to tailor the translation function to the needs of an organization or a specific group of users in an organization. One example is an administrative office where staff handle general correspondence and order information. A user can process these documents within the Google Docs environment and "move" the translated text into another system by clicking an icon to activate the developer's script. With more effort, Google Docs provides templates and tools to create an application to handle creating invoices in English and a target language. Weaving together Google functions continues to become quicker and easier with each incremental tweak to the Google Docs and Google Apps ecosystem.

The established vendors of translation programs will have to raise the level of their game. Google's basic line up of more than 50 language pairs and its competitive pricing translate to increased competition. Large-scale online translation systems can hit six figures when deployed across an organization. Google's approach can drop that cost to a fraction of the established vendors' fees.

Think of Google’s translation and Apps functions as your organization’s language *wallah*.

Stephen Arnold, ArnoldIT.com

www.arnoldit.com/sitemap.html

www.arnoldit.com/wordpress

Mr. Arnold is an independent consultant residing in Harrod's Creek, Kentucky. He is the author of more than 10 monographs. His most recent is Google: The Digital Gutenberg, available from www.infonortics.com. He writes monthly columns for KMWorld, Information World Review, and the Smart Business Network.