

Boosting Performance of 3rd Party MT Products

By Vadim Berman



(This document assumes that the reader is acquainted with the basic concepts of machine translation - MT.)

Carabao Language Kit is a family of linguistic products developed by Digital Sonata. With the significant costs and efforts invested in machine translation packages, it is no surprise that the customers are reluctant even to consider migrating to other software, no matter how promising it looks.

While Carabao has a native Translation Server component, we also offer tools aimed to enhance and improve the performance of existing solutions.

Using Carabao DeepAnalyzer to Tune-up Accuracy

The first and the best known example of high-quality automatic translation is METEO, a Canadian system built to translate weather forecast. The secret was simple: it was limited only to one domain of discourse. MT systems evolved to contain specific dictionaries for specific domains of discourse; 90% of the modern solutions contain some sort of domain classification. Some front-ends include a combo box, which allows the user to select a domain of discourse. This is supposed to improve the accuracy. However, there are two problems with this approach:

1. What if the user did not select the domain correctly, or did not select it at all?
2. What if the text contains more than one domain of discourse which differ from sentence to sentence?

If some magical piece of software was able to determine the domain of discourse on a sentence level, these problems could be magically solved. Statistical subject classifiers do not work here – they require bulk amounts of text and do not provide sentence level feedback.

Fortunately, now there is such software. Meet Carabao DeepAnalyzer, a component which does exactly that (among other features): analyzes a text and returns a set of domains in use, breaking them down to sentence level.

The principle is the following:

1. The content to be translated is fed into Carabao DeepAnalyzer, which maps the domains of discourse.
2. The 3rd party machine translation provider sets as a parameter domains or subject areas matching the domains returned by DeepAnalyzer, per sentence. As there might be more than one domain, the first matching one must be returned.
3. The 3rd party machine translation provider processes the content.

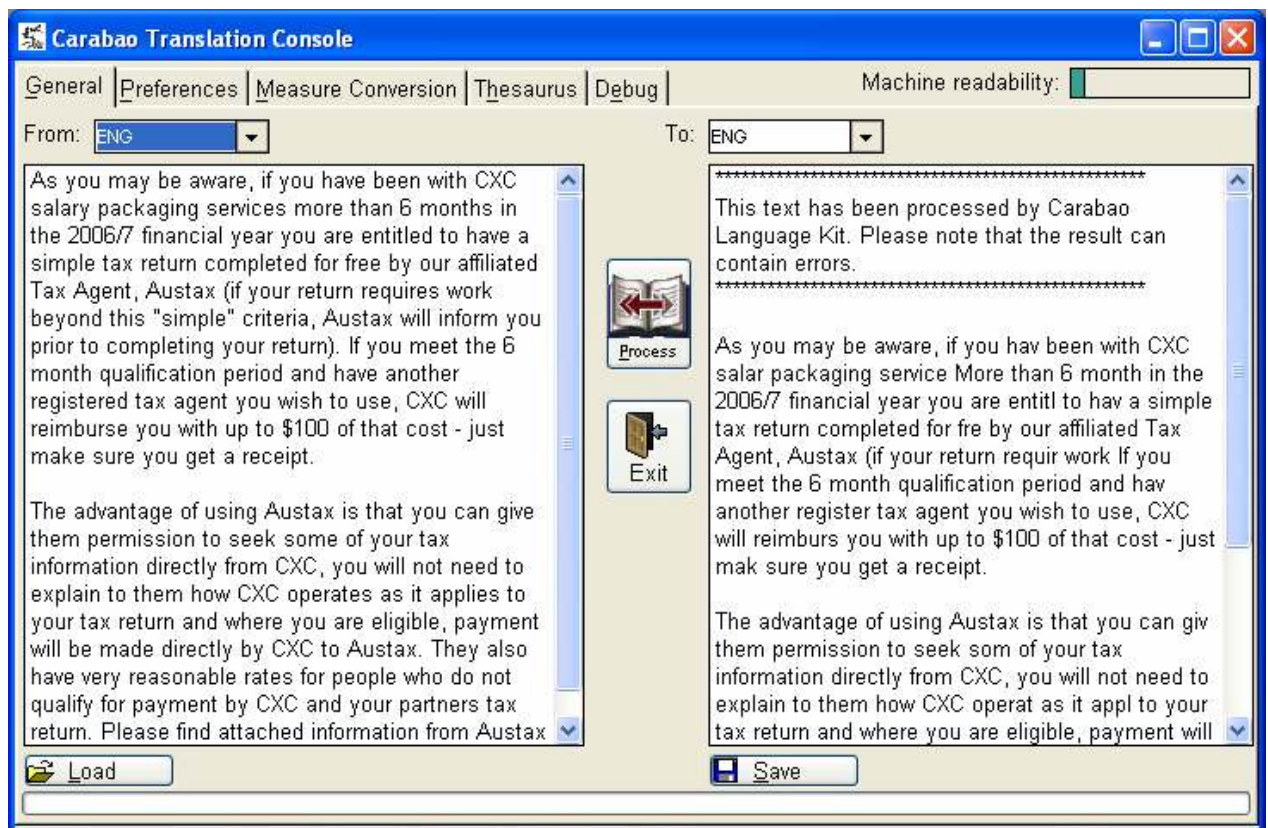
Carabao does not have a set of fixed domains. Instead, every concept in the dictionary is referred as a potential domain. Therefore, when mapping Carabao's domains to your software's domains, there is much flexibility. For example, if there is no specific "banking" domain in your software, but there is "finance", you can map Carabao's "banking" to your "finance" dictionary.

Try before You Buy

There is no silver bullet; different packages have different strengths and weaknesses, and therefore the degree of improvement from implementation of *Carabao* components also varies.

The distribution of domains can be quite different for different customers. With licensing and development costs involved, we made sure it is possible to run a quick check in order to approximate the performance yield. See below the step-by-step instruction on testing Carabao domain extraction with IBM WebSphere Translation Server. We would like to reiterate though, that **the principle works for any MT engine that supports selection of domains / subjects**.

1. Go to www.digitalsonata.com/download.aspx?type=desktop , download and install the free Carabao Standard Edition. Do not worry about nag screens or expiration, this desktop edition is completely free.
2. Obtain 6-8 examples of the input content.
3. In Carabao Translation Console, set English as both input and output languages. Feed your example as shown on the figure below:



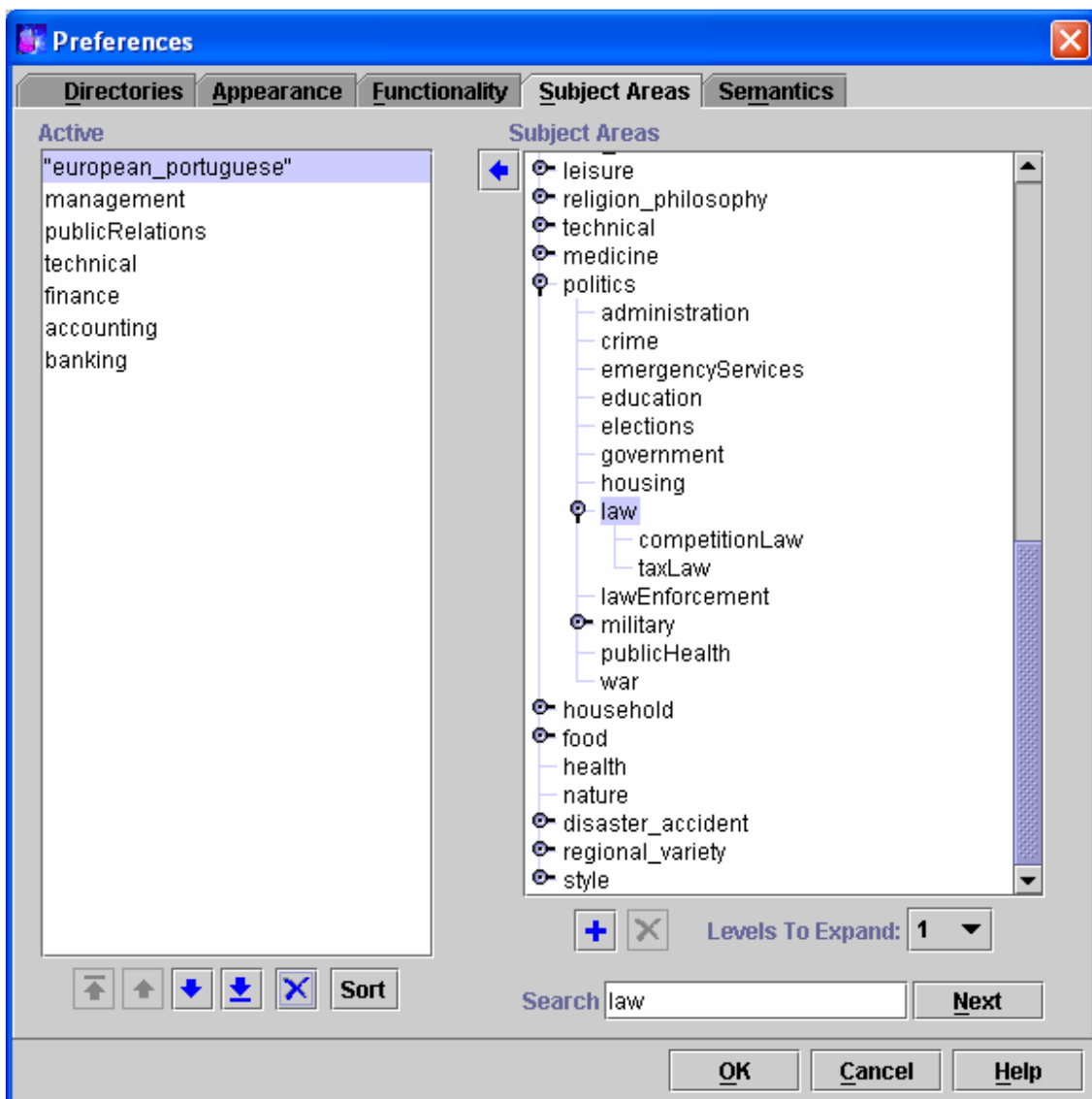
4. Once the processing is done, click on the Debug tab. Inspect the rightmost table which contains the domains:

Show tracer window

Range	Domain	Range
1 - 4	linguistic communication	1 - 5
1 - 3	money	1 - 4
1 - 3	law	2 - 3
1 - 3	revenue enhancement	1 - 5
1 - 2	written communication	1 - 3

5. Take note of the domains that can be mapped in IBM WebSphere Translation Server, and their ranges. In our example we will map:
 - a. Sentences 2 and 3 to WTS subject areas "law" and "finance"
 - b. Sentences 1, 4 and 5 to WTS subject area "finance"

In WTS, you might select a custom subject area more suitable for your customer. Go to WTS User Dictionary Manager, open Preferences and this is where you can find the subject areas:



6. In WTS User Dictionary Manager, use the *Test Translation* option (or any other client that has subject areas enabled) to check how the sentences will translate when the subject areas are selected. Enter each sentence separately setting the subject areas as mapped:
 - a. For sentence 1, set the subject areas to “finance”
 - b. For sentences 2 and 3, set the subject areas to “finance” and “law”
 - c. For sentences 4 and 5, again set the subject areas to “finance” and “law”.
7. Repeat steps 3 through 6 with the other fragments.

Note that the process of mapping is not as difficult as it might look. The dictionary entries in Carabao are interconnected with links such as hypernym (generalization) or domain. In case of a very specific domain such as “revenue enhancement” above, a better bet is to use the API of Carabao DeepAnalyzer in order to discover what domains this domain is a part of (if any), or what domains do its ancestors belong to.

It is important to remember to feed an entire fragment. Carabao is capable to follow the context and the result of analysis can be quite different if the sentences are fed one-by-one.

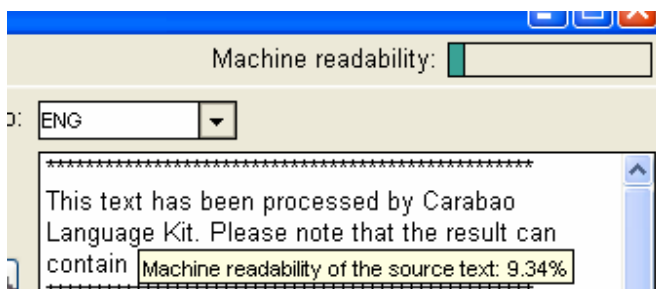
Is the result always 100% accurate? While you can see that Carabao is extremely accurate, it cannot be perfect, just like no natural language processing software is perfect. It is natural to ask, is it possible to know when we can trust Carabao’s analysis? This brings us to our next item –

Translation Workflows and Machine Readability

As mentioned above, no natural language software is perfect. In many cases, there is a need to “fall back” to the old and (more) reliable human translation. But how do we know whether a text was too difficult for a machine to handle?

So-called readability indices have long history, yet they are useless in our case. Usually long, rare words more difficult for humans (for example, particle accelerator or plutonium) are more machine-readable, while short, ambiguous words which humans decipher with little effort, are guaranteed to be misinterpreted.

Therefore, we designed Carabao Readability Indicator, a component aimed to measure machine readability. In the desktop suites, it is exposed via a bar graph over the output field, and via a tool-tip shown when the mouse hovers over the output field:



The higher the number, the easier the text for natural language software in general. Of course, different packages perform in a different way, and all have strengths and weaknesses – just like different people have different strengths and weaknesses of their reading skills. The result simply indicates how difficult a text is for a machine to process and understand correctly. Before use, a “pass” threshold must be evaluated by a few samples. The practice shows that in case of Carabao, texts with machine readability above 15% are easily analyzed; those between 5% and 15% can be relied on in most cases, and those with readability below 5% are hit and miss.

The component is invaluable in translation workflows, where it is necessary to know what texts can be translated by machine translation and what texts must be left to humans.

Summary

With the narrow choice of natural language processing products, complementary products and low-level middleware such as *Carabao* components will encourage further growth by making development easier.

The principle described in the document can be applied to any other application working with unstructured natural language data and having “doubts” about candidate words; the first ones to try are OCR and speech recognition components. It can be also used to implement smarter auto-complete GUI.

The technology implemented in *Carabao* family of products presents new and exciting opportunities. We are merely scratching the surface in this document; there is much more to *Carabao* than what we mentioned in this document. Please do not hesitate to inquire should you need any clarification on this or other topics.

Contact Details



Vadim Berman
vadim.berman@digitalsonata.com