Background Search for Terminology in STAR MT Translate

Giorgio Bernardinello
STAR Group - Wiesholz 35
8262 Ramsen
Switzerland
giorgio.bernardinello@star-group.net

Judith Klein
STAR Group - Wiesholz 35
8262 Ramsen
Switzerland
judith.klein@star-group.net

Abstract

When interested in an internal web application for machine translation (MT), corporate customers always ask how reliable terminology will be in their translations. Coherent vocabulary is crucial in many aspects of corporate translations, such as documentation or marketing. The main goal every MT provider would like to achieve is to fully integrate the customer's terminology into the model, so that the result does not need to be edited, but this is still not always guaranteed. Besides, a web application like STAR MT Translate allows our customers to use – integrated within the same page – different generic MT providers which were not trained with customer-specific data. So, as a pragmatic approach, we decided to increase the level of integration between WebTerm[1] and STAR MT Translate, adding to the latter more terminological information, with which the user can post-edit the translation if needed.

1 STAR MT Translate

STAR MT Translate is a highly customisable web application for machine translation (MT). It is not designed to be part of an automated translation process, nor to be a tool for expert translators, but rather to help any employee of a company understand texts and documents written in foreign languages. The UI can be designed to fit the corporate style of the client and it offers easy access to the STAR MT engines, specifically trained for each customer, as well as for the most well-known online MT providers.

2 TermAssist

In the last few years, we have seen many companies starting to offer connections between terminology and MT, like the dictionary in GoogleTranslate or the glossary in DeepL. STAR started working almost five years ago on an integrated solution, where corporate terminology can be retrieved from the same webpage in which a text has been translated using MT.[2] The purpose was to give the MT user the possibility of consulting the company's dictionary without switching tabs in the browser, by simply highlighting one or more words. This function was named TermAssist and has since become one of the most requested functions of STAR MT Translate. The main limitation of this kind of approach is the lack of matches for inflected words. A very detailed dictionary could also contain inflections referring to the main term, but this is not realistic in practice, and it may get complicated for multi-word concepts. For example, the plural of the German word "*Sitzplatz*" (seat) is "*Sitzplätze*" (different vowel inside the word plus -e added at the end) while the corresponding Italian forms are "*posto a sedere*", singular, and "*posti a sedere*", plural (the last letter of the first word contains the inflection).

3 Background search

The solution for such cases comes from a further implementation of STAR terminology

[1] WebTerm is the STAR web-based terminology application.

[2] Bernardinello, G. 2018. Terminology validation for MT output. EAMT 2018, 21st Annual Conference of the European Association for Machine Translation., p.343, Alicante, Spain

applications: the background search. Already used in Transit during the import phase to look for all available terms in the TermStar dictionaries, the function was adapted to become a REST API extension.

Thanks to the integration of the background search in STAR MT Translate, when a user translates a sentence, regardless of whether its result came from MT or translation memory[3], all terminology matches are shown[4] in both the source and target languages annotating the text without modifying it. The background search is more accurate than the previous TermAssist search and it is able to find inflected forms in the most common languages. This solution is more efficient even at first glance, since the user can immediately identify all concepts with a terminology entry instead of manually highlighting text and looking for more information. Furthermore, the algorithm checks whether the concepts found in the source correspond to the ones found in the target. E.g.: when the user points the mouse at a term in the target language, both the term itself and its corresponding form in the source language, if available, are highlighted. The same works, of course, the other way around. This can be very helpful when the user is not familiar with one of the two languages; in fact, he or she can verify with a click if the automatic translation is handling that specific concept with the desired corporate-specific vocabulary. If not, the user can view a list of allowed synonyms and related words by right-clicking on the concept.

4 Negative terms

Negative terms, or disallowed terms, are possible translations of a concept which are either wrong or not accepted by the language department of a company. In both cases, the ability to identify them quickly in an automatic translation represents another pertinent advantage for the end user. It happens quite often that the customer needs to specify some negative terms which are only disallowed for that specific concept, but they may be correct in other contexts. For example, a company may want to avoid a colloquial form like "*car*" when translating from the German "*Fahrzeug*" (vehicle), but "*car*" may be accepted when translating the more informal "*Auto*". Thanks to the double-check between source and target texts, the application can highlight a term differently depending on its counterpart; this will give the user visual input on critical words with the possibility to change them to a valid synonym.

This aspect is crucial for customers using different translation providers, since their translation may not always match the desires of the company regarding terminology; even a user with no experience or terminological expertise can immediately see if the text contains invalid concepts, correct them, and continue working with a translation more consistent with the desired corporate language.

5 Future developments

An interesting extension of this feature can be achieved with the contribution of TMC (Translation Memory Container), the STAR database for reference material. It is already possible to activate the connection to the database in order to retrieve translations from the TMC in case of perfect matches. This will skip MT for texts which have already been translated and approved by the company.

The TMC could also be used together with the background search to retrieve segment pairs which contain the same terms found by the background search in both the source and target language. Transit already uses this context-specific TMC search as a support for professional translators who can then see other examples of complete sentences where specific terminology has been used. This is particularly important when a language has more possible translations for the same term in another language. As an example, we can take the English word "*glass*" meaning an object we use to drink from or the material from which it is made. In Italian there is only one possible word for the object, "*bicchiere*", and one for the material, "*vetro*". In cases like this, where both translations are valid, the user can find help in some context-based translations that the company has already completed and approved. This function could be a significant addition to the information given by the dictionary, which may contain some useful examples, but not always one for each specific form of the term.

[3] STAR MT Translate offers the possibility to search for the sentence in the reference material and only send it to MT if not found.

[4] Each customer can decide how to differentiate terminology matches from normal translated text. It can be any CSS property: different colour, underline, different font, etc.