MTradumàtica: Free Statistical Machine Translation Customisation for Translators

Gökhan Doğru
Universitat Autònoma de Barcelona
gokhan.dogru@e-campus.uab.cat

Adrià Martín-Mor
Universitat Autònoma de Barcelona
adria.martin@uab.cat

Sergio Ortiz-Rojas
Prompsit Language Engineering
sergio@prompsit.com

Abstract

MTradumàtica is a free, Moses-based web platform for training and using statistical machine translation systems with a user-friendly graphical interface. Its goal is to offer translators a free tool to customise their own statistical machine translation engines and enhance their productivity. In this paper, we aim to describe the features of MTradumàtica and its advantages for translators by focusing on its current capabilities and limitations from a user perspective.[1]

1. Introduction

The working environment of modern translators has been changing drastically. While there are still some translators trying to adapt to the advent of computer-aided translation (CAT) tools, now there is a need to adapt to a new working environment which also includes machine translation (MT). However, MT systems are presented generally as black-box solutions in which translators cannot intervene, make modifications or customisations. Hence, the translators are dependent on MT solutions provided by either their language service providers or huge corporations.

We see the availability of free statistical machine translation (SMT) systems like Moses as a unique opportunity to narrow the technology gap between human translators and MT technology, and therefore to increase the effective usage of this technology. For the last few years, building and training SMT systems by end users has been a complex task involving a number of computing skills which might prevent the adoption of the technology. Therefore, we think that whenever necessary tools (free, open and easy-to-use tools) are presented to the translators, this gap can be eliminated to some extent, and translators can be empowered and be prepared to be competitive in the sector. With this assumption in mind, we have developed a Moses-based web platform, MTradumàtica, within the scope of ProjecTA.

ProjecTA (www.projecta.tradumatica.net) is a Tradumàtica group research project (www.tradumatica.net) at the Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental at the Universitat Autònoma de Barcelona. It works from the basic assumption that translators have the appropriate profile to manage MT-related tasks, and that empowering translators in MT tasks is beneficial for

[1] This work was supported by the ProjecTA project, grant number FFI2013-46041-R [MINECO / FEDER, UE].

translation companies. The project was split in two phases: first, to explore how MT is used by the translation sector in Catalonia and Spain through a survey sent to 187 translation agencies; second, based on the survey responses, to develop software to bring MT about closer to translators. The conclusion of the first phase is that MT use among most translation companies in Catalonia and Spain is low. Hence, ProjecTA decided to focus the second phase in the development of a software that can eliminate some of the barriers to implementing MT systems in the translation industry. These considerations have led to the creation of MTradumàtica.

2. What is MTradumàtica?

MTradumàtica is a free, Moses-based web platform for training and using SMT systems with a graphical user interface. Users can create their own engine in a few steps by uploading sentence-aligned parallel files in the usual Moses text format, then use these files to train a translation model and a language model, and ultimately train an SMT engine. To put it simply, there are 5 steps: (1) Upload files (2) Create and manage monotexts (3) Build language models (LMs) (4) Create and manage bitexts (5) Train SMT translation models. The LM, in the context of SMT, is the statistical model of a natural language, while the translation model (TM) includes the translation probabilities derived from parallel corpora. Monotexts are the monolingual texts used to create the language model, while bitexts are aligned bilingual texts (for example, a technical text aligned sentence by sentence with its translation) used to create the translation model. These two types of texts provide the training data for SMT to operate on.

At the end of the training, users can use their engine to translate texts or documents within the website. This means that translators can use their own resources or open resources (such as corpora from the Opus collection http://opus.lingfil.uu.se) and customise their own engines according to their needs. As stated above, MTradumàtica aims at empowering translators in the context of the local translation fabric, made up mainly of small companies. Although the corporate perspective typically confines translators to mere end-users of MT, MTradumàtica aims at allowing them to develop their own engines and use them within their own personal, low-scale workflows.

The current version of MTradumàtica is available from GitHub (http://github.com/tradumatica/mtradumatica). It comes with a semi-automated installation procedure that works on Linux (local and server) and relies on technologies such as Python and Docker, as well as the software usually coming along with the Moses SMT system and other pieces of software from the Apertium project.

3. The Advantages and Limitations of MTradumàtica for Translators

One of the assumptions of SMT is that building MT engines from domain-specific parallel corpora tends to increase the quality of the raw output and, therefore, productivity. Considering that professional translators generally work on specialised domains for long time and collect huge amount of parallel corpora in time (under the form of translation memories), they can build their own engines and use them on a project-based basis. Since this customisation is made on the web platform, translators can use any operating system, provided that it has a web browser and an internet connection. However, there are still some developments needed for MTradumàtica to be fully functional for translators. Considering that most translators work with computer-aided translation (CAT) tools, their parallel corpora are generally exported in Translation Memory Exchange (TMX) format. Nevertheless, in the current version, it is not possible to upload TMX files to the file manager of MTradumàtica. Despite the fact that converting TMX to a moses file format is an easy task, the addition of the TMX upload feature will make MTradumàtica more convenient for translators. Secondly, for the

same reason, the integration of MTradumàtica with CAT tools through an API key will allow the translators to use their SMT engine within their own work environment. Thirdly, automatic evaluation metrics such as BLEU are needed to be able to evaluate the quality of the SMT engine beforehand so as to decide whether its quality is high enough to be used for translation tasks. Fourthly, confidentiality is a very important issue for translators (since they enter into non-disclosure contracts with their clients). The platform shall provide private user space (an account with a username and password) and guarantee that the parallel corpora are not used by anyone else. Although these are the prioritised features from the point of view of translators, some other features such as concatenating and prioritising models through GUI, terminology management, integrated corpora management, automated pre and post-editing functionalities shall be added to MTradumàtica. Currently, a feature called *Inspect* is also available. This feature, partially functioning at the moment for demonstration purposes, should allow the user to query and examine the components of the engines already created, i.e., the TM and the LM.

4. Concluding remarks

This paper has shortly described the current state of the MTradumàtica platform and its further developments. There is a certain need for a free machine translation platform for translators to remain competitive in the translation sector. MTradumàtica attempts to ease integration with the workflow and to remove most of the technical barriers for the integration of MT in enterprises so that freelance translators and small companies can use it. MTradumàtica is available at the moment for testing purposes at www.m.tradumatica.net.