Project Adaptation for SMT Over Several Days.

Frédéric Blain, Amir Hazem, Fethi Bougares, Loïc Barrault, Holger Schwenk
LIUM, University of Le Mans

Abstract

Computer assisted translation (CAT) tools, which are designed to support and facilitate the translation process play an important role to boost the productivity of professional translators. However, and despite significant improvements, CAT tools are not yet completely satisfactory. One of the reasons is that most MT systems are not fully integrated with the human translation workflow. CAT tools are usually based on translation memories (TM), terminology dictionaries, concordancers, spell checkers and recently machine translation (MT) systems. If a segment to be translated is present in the TM, the CAT tool displays the corresponding translation through its editor so that the translator can use it directly, on the contrary, if a segment to be translated is not present in the TM, then the CAT tool uses the MT system to translate the segment. If an MT system is not yet able to produce a suitable translation (which is comparable to a translation produced by a human translator), at least it can help translators to gain time while they just need to correct the MT output instead of translating the whole segment (sentence, paragraphe or document).

A classical translation scenario consists of translating a set of documents by human translators over several days assisted by a given CAT tool. After the first working day, the translated texts and user corrections could be injected into the MT system to allow the system to adapt and to learn from its errors. We call this process project adaptation (PA). It can be repeated throughout the duration of the translation project (several days). The main goal here is to make the MT system more specific to the project and fully integrated with the human translation workflow, in order to minimize the MT output errors throughout the duration of the project and to reduce as much as possible human intervention.

If it has been already shown that post-editing MT output increases the productivity of human translators (Guerberof 2009, Plit and Masselot 2010, Federico et al. 2012, Green et al. 2013), specializing the MT system on the documents to be translated is relatively new (Mauro et al, 2014), and to our knowledge there is no reported work on such a case study over 5 days of work. In this paper, we report recent investigations from the European MateCat project which consists of a Web-based CAT tool for professional translators that allows self-tuning MT based on domain and project adaptation technique. Along with project adaptation, domain adaptation (DA) only allows an MT system to be specific to a particular domain but not to a particular project. The purpose of this study is to show that corrections performed by translators over one day of work are an important and a valuable resource to improve the MT system for the next day. We investigate how a domain specific statistical MT system can adapt day-to-day from collected translators feedback on day one to improve the performance of the system on day two and so on. This is done by performing data selection (Moore and Lewis 2010, Axelrod et al. 2011) after each day of work based on translators feedback and retraining the MT system using the new selected data for the next day. We conduct experiments on the English-French language pair on the legal domain over 5 days and show the significant improvements when project adaptation is performed.

2nd Translation in Transition Conference. Mainz, Germany, January 2015. To appear