Multilingual Access to Educational Material Through Contributive Post-editing of MT Pre-translations by Foreign Students

Ruslan Kalitvianski[1,2(✉)], Valérie Bellynck[2], and Christian Boitet[2]

[1] Viseo Technologies, 4 avenue Doyen Louis Weil, Grenoble, France
ruslan.kalitvianski@imag.fr
[2] LIG-GETALP, Bat. IMAG B, 41 rue des Mathématiques, 38400 Saint Martin d'Hères, France
{valerie.bellynck,christian.boitet}@imag.fr

Abstract. In our teaching practice, we often observe that, due to the lack of prerequisites and limited mastery of a language, foreign students face difficulties in understanding course contents. This especially burdens students from Eastern and South-Eastern Asia, because of the distance between their native languages and the instructional language (French in our case). We propose a quick and cost-effective method for making educational content accessible in the native tongues of the students, through a contributive computer-assisted multilingualization by voluntary participants. The process consists in post-editing MT (Machine Translation) pre-translations via an interactive multilingual access gateway (iMAG), which displays a web page in a selected language. Since 2012, several students have validated the approach by producing in Chinese more than 500 pages (125 K words) of French undergraduate and graduate course material about computer science, at a rate of about 10 min (total time) per standard page. This multilingual resource is freely accessible on the MACAU-Chamilo platform.

Keywords: Multilingual access · Educational material · Computer-assisted translation · Post-editing

1 Introduction

Our university receives each year about 2300 foreign students. Around 650 of our own students spend a part of their studies abroad via student exchange programmes such as Erasmus[1].

Their academic success depends heavily on their mastery of the instructional tongue, in our case French, and, to a lesser extent, English. But, unfortunately, their linguistic skills are often too limited. It is also common for such students to lack scientific prerequisites necessary to follow the courses of the host university, and as a consequence they have to spend additional time acquiring them.

[1] http://ec.europa.eu/programmes/erasmus-plus/index_en.htm.

F.W.B. Li et al. (Eds.): ICWL 2015, LNCS 9412, pp. 109–118, 2015.
DOI: 10.1007/978-3-319-25515-6_10

When faced with difficulties in French, some seek books in English, however they encounter two important problems:

– these books are barely helpful to them if their English skills are no better than their French, a frequent case with our students from East Asia
– the notations in these books often differ from what is taught in our classes, and they don't cover the same topics, with the same level of detail.

Thus, these students need to get access to course material in the tongue they know best, and in sync with what is taught in our university.

Motivated by this observation, we started the MACAU project in 2012, aiming at providing a multilingual access to the educational content produced by professors, lecturers, as well as students, such as books, hand-outs, lecture notes, report papers, exam papers, solutions to exercises, etc.

A naive approach to multilingual access would be to use a free online machine translation (MT) service, such as Google Translate[2] (GT), without any further ado. GT offers a wide choice of language pairs, however it presents important problems.

1. The quality of translations, though quite acceptable for short conversational sentences, deteriorates for narrowly specialized and advanced technical areas that are taught in our university, as well as for many language pairs, and for longer sentences.
2. While GT allows suggesting corrections to translations, these corrections are not displayed upon subsequent visits to the page. They are stored in Google's translation memory and used for retraining later its statistical MT system.
3. GT requires a URL to a file repository where course material would be stored.

Although rough machine translations are of limited usefulness for multilingual access to educational material, they can be very helpful for accelerating human translation [1]. The approach adopted in the MACAU project described in this paper is to immediately give access to the pedagogical material (formatted in html) in the desired access language, using MT "pre-translations", and then to improve the quality of target segments in an incremental and contributive fashion. In other words, if a student is not satisfied with the proposed translation, s/he can correct it directly on the web page, in a seamless manner.

The rest of the paper is organized as follows: in the next sections, we discuss previous and similar work, then describe the platform and its features, and lastly discuss the resources obtained as by-products, as well as the encountered difficulties.

2 Prior Work

There seems to be not much prior work. Two projects that stand out are the EU Bologna project and the more recent SlideWiki contributive project.

[2] https://translate.google.fr/.

2.1 The Bologna Project

The Bologna project[3] was a EU-funded initiative aiming at building "a translation service designed for translation of course syllabi and study programs from 9 languages — Dutch, English, Finnish, French, German, Portuguese, Spanish, Swedish and Turkish — into English", using computer-assisted translation tools. Chinese was later added as Chinese students often outnumber all other nationalities among foreign students. This three-year project ended in 2013, and offered a demonstrator of the collaborative web platform, however it has not led to a permanent web service. The translation tools were to be specifically adapted to the translation of course syllabi. In 2013, we evaluated the online Bologna demonstrator and found its translations to be of a quality inferior to that of GT. This service has since been discontinued.

Several ideas underlying the Bologna project converge with those of our project:

- collaborative approach to translation and its improvement
- usage of translation memories specialized to each context
- definition of roles and tasks, such as translator, post-editor, moderator, MT developer, etc.
- handling different formats (html, docx, xslx, txt, rtf, URL link).

However, Bologna had both conceptual and implementation flaws.

- The project lacked ambition, as it was limited to translating 9 of 22[4] European languages into English and Chinese. International students arriving in a foreign country are mostly non-native speakers of English and do not have a sufficient mastery of English to really understand translated documents, and even less to contribute to the improvement of machine pre-translations, as one should always post-edit into one's native tongue.
- The MT systems that were demonstrated produced output of inferior quality. This is due to the use of statistical MT, which can produce useful results only if it is trained on a large or very large corpus of parallel translations of good quality.
- Access to the post-edition interface was restricted to approved users, and the interface itself was cumbersome. In order to elicit contributions, the interface should allow for post-edition directly on the displayed document, and be freely accessible.

2.2 SlideWiki

SlideWiki is a recent project aiming at the online collaborative construction of educational presentations [2]. These presentations can either be built on the website, or be imported from a pptx format. An interesting aspect is the possibility of producing versions in a different language using Google Translate. However, limiting the content type to presentations appears too restrictive, and the lack of a translation memory limits the efficiency of the translation process.

[3] http://www.bologna-translation.eu/.

[4] There are now 23 official languages in the EU, but Croatian was added only in 2013.

2.3 Interactive Multilingual Access Gateways

The concept of an interactive multilingual access gateway (iMAG) has been proposed by Boitet and Bellynck in 2006 and has been used in our laboratory since November 2008 [3]. An iMAG is a gateway very much like Google Translate at first sight: one specifies the URL of a web page and the access language, and then navigates in that access language. The iMAG displays the translated web page with the layout preserved. When the cursor hovers over a segment (usually a sentence or a title), a palette displays the source segment and proposes to contribute by correcting the target segment, in effect post-editing an MT result.

Contrary to GT, an iMAG is dedicated to an elected Web site, or rather to the elected sublanguage defined by one or more URLs and their textual content. It contains a dedicated translation memory (TM). Segments are pre-translated not by a unique MT system, but by a (selectable) set of MT systems. Systran and Google Translate are mainly used now, but specialized systems developed from the post-edited part of the TM have also been used, notably for French→Chinese (Fig. 1).

Fig. 1. The post-editing palette on a segment.

While reading a translated page, it is possible not only to contribute to the segment under the cursor, but also to seamlessly switch to an advanced online post-editing environment, equipped with proactive dictionary help as well as filtering and search-and-replace functions, and then return to the reading context.

An MT middleware, TRADOH, allows us to select, parameterize and call the MT systems and define the translation "routes" used for various language pairs. An iMAG-relay is planned to manage users, groups, projects (some contributions may be organized, other opportunistic), and access rights. But, for the moment, these functions are managed

by the "back-end" corpus and TM manager, SECTra_w. MT systems tailored to the selected sublanguage can be built and have been built (by combinations of empirical and expert methods) from the TM dedicated to a given elected Web site. That approach inherently raises the linguistic and terminological quality of the MT results, that can sometimes produce *raw* rather than *rough* translations.

Besacier [4] reported in 2014 on an experiment on collaboratively translating into French a short English language novel via an iMAG. In this experiment, which involved non-professional translators, he showed that costless translations of literary texts of acceptable quality can be produced relatively rapidly by post-editing volunteers, even though such translations initially present a certain lack of unity, as well as stylistic inadequacies typical of a beginner translator's work. For our purposes, however, stylistic considerations are less relevant.

3 Proposed Solution

Our aim is to provide a platform allowing users to upload their documents, and to access these documents in the languages of their choice. The translated versions should preserve the layout of the original document, as well as allow users to edit the translations where needed, collaboratively and incrementally, through direct interaction with the concerned segments.

Language learners find it helpful to be able to see both the original text and its translation simultaneously. It allows them to learn sentence to sentence correspondences between languages. We therefore should provide a means to display in parallel the source text and the translation.

Although the access to the content should be open to all, some rights management policy for post-editing should be implemented. The iMAGs can be configured with several modes of control or "moderation", somewhat like Wikipedia.

To support the pedagogical documents, we use an open-source e-learning platform Chamilo, instances of which are widely used by our universities. It features a multilingual interface, and allows users to create courses either by uploading existing HTML documents or by building them online via a WYSIWYG HTML editor. It also allows to communicate via forums or instant messages, as well as to define dictionaries.

We have thus set up a Chamilo platform and equipped it with tools for selecting the access language of a document. The list of languages is defined by the available MT systems or translation memories (Fig. 2).

The default access language of a document is its original language. To access it in another language, one selects it in the "AXiMAG" menu and clicks on "Translate". The resulting course page is reconstituted from the MT results or from the available post-editions, which are both stored in a TM (translation memory) managed by the MACAU iMAG. This works for any HTML content available on MACAU-Chamilo, whether it has been created through the tools of the platform or uploaded by a user. It has to be noted that we are not restricted to Chamilo: iMAGs can be easily integrated with any other platform that provides a URL to its course material.

Fig. 2. Multilingual access integrated into Chamilo.

For a course that has not yet been post-edited, the first translation is obtained by MT. The user can correct the translation via the palette that appears when the cursor hovers over the sentences, and these corrections are saved in the translation memory managed by the SECTra_w "backend" of the iMAG. The (system-assigned) reliability level and the (user-assigned) quality score are used for ranking the translations and post-editions in the memory; the post-edition with the highest score is displayed during the visit of the page via the iMAG.

The correction process is called "post-editing", as opposed to "revising". The difference is that it is absolutely necessary to read and understand every sentence before correcting the "pre-translation". This is why we regularly ask good foreign students in the classes where we teach (undergraduates and graduates in computer science) to make the first post-editions.

The user can switch between the parallel view, which displays both the translation and the original, and the translation view, which displays only the translation. The optional "reliability brackets" around segments allow to see at a glance which have been post-edited: green brackets indicate that the segment has been validated by a moderator, yellow brackets that it has been post-edited but awaits moderation (for contributions by users that are not registered), and red brackets enclose MT results (Fig. 3).

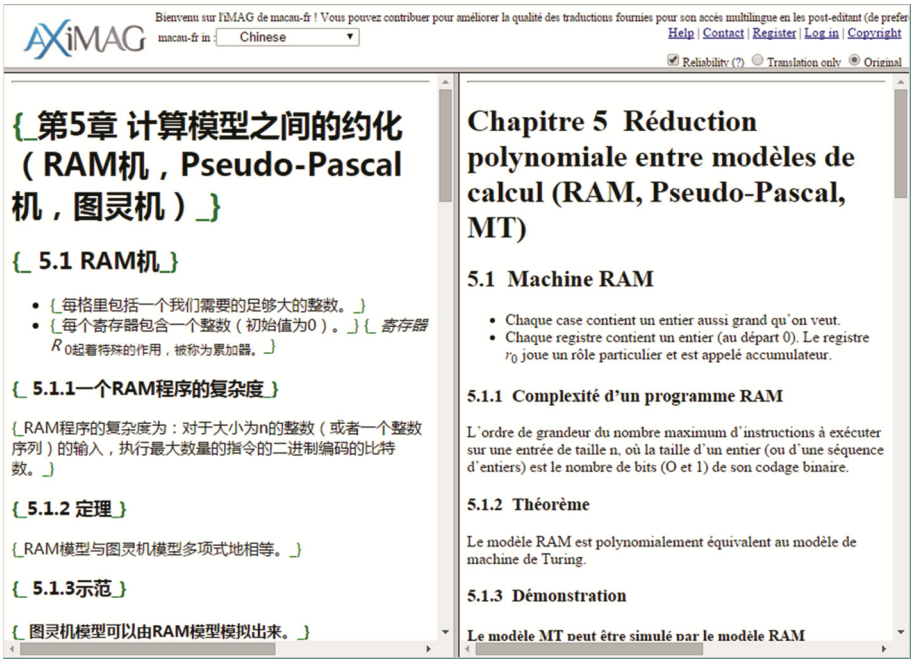

Fig. 3. A parallel presentation of a document, in target and source form, displaying optional reliability brackets around target segments.

4 Method and Results

Our experiments so far confirm the hypothesis that post-editing machine translations of course material by volunteers is a viable way of producing versions of "sufficient quality", even when the MT systems used are not of good quality, if judged by translators. Usage quality does not necessarily correlate with linguistic quality.

The first step is to collect material and convert it into HTML. We have collected educational documents about computer science produced by our teachers and students. These documents include a book on logic ("Logic and automatic demonstration" by S. Devismes, P. Lafourcade, and M. Lévy), lecture notes on computational complexity, as well as various handouts.

Documents came in different formats. The book and lecture notes were in LaTeX and had to be converted into HTML via tools such as HeVeA[5] and LaTeX2HTML[6]. Others were in Microsoft DOCX format, an XML-based format whose conversion into HTML is straightforward and performed well by office suites such as Microsoft Office, LibreOffice and AbiWord.

[5] http://hevea.inria.fr.
[6] http://www.latex2html.org.

The situation was less favorable for PDF files. When this experiment was conducted, there were no tools of acceptable quality for converting PDF into HTML. The available tools either only extracted the text, disregarding the document layout, or produced HTML documents that attempted to preserve the typographical layout of the pages, but in doing so produced HTML code quite difficult to parse for submitting to MT systems. Progress has since been made by Microsoft Word, which now can transform some PDF files into Word documents. This is fortunate, as some teachers are only able to supply us with PDF files, and not their sources.

A second step is the segmentation into pages of convenient size for the MT system we used (GT in this experiment), typically the size of a chapter. This step was done automatically via SegDoc, a segmentation tool for potentially marked-up text that we developed for the purpose.

A further crucial step is normalization, which consists in selecting sections of HTML that should be protected from translation, typically mathematical formulas in their alpha-numeric transcription, as well as algorithm code, both susceptible to be treated as text by MT systems. For instance, a variable named '`I`' may be interpreted as a first person pronoun, which is problematic. Other literals may be removed or inverted by the MT system, thus deforming the entity. Protecting these sections consists in inserting the attribute "`translate=no`", part of the HTML5 standard, into surrounding HTML tags.

As for the moment we have no automatic tool for detecting these non-linguistic fragments in most cases, this step had to be done manually. One perspective of this work would be to employ a classifier for automatic detection of such entities.

Once the documents were prepared and uploaded, we incited foreign students (mostly Chinese) to perform some post-editing. As a result, 70 HTML documents, totaling 16069 segments (sentences or titles) of an average length of 8 words per segment, have been post-edited into Chinese. This represents about 514 standard pages. The Table 1 shows the current status of the platform.

All these documents are freely accessible to the public[7]. We are also open to creating new iMAGs for those who are interested in our approach.

A detail worth noting is that post-editing course material into Chinese proved to be quite useful for some students, as it helped them prepare some exams requiring a good and fast understanding of quite lengthy and complex exam papers (in French). A striking example of this is a student who earned excellent marks during the semester in situations where he was able to read the instructions at his own pace and wasn't required to produce lengthy explanations in French on the spot, and yet scored 2.5/20 at the final exam. After post-editing with us, his grade rose to 11/20 and he passed. We conclude that this process helped him make progress both in the subject domain and in expression in French.

With this experiment, going on since 2 years, we have thus

- shown that our approach is a viable way of producing multilingual content from monolingual sources
- produced a significant amount of documents in Chinese, with free access
- seen that post-editing helps understand the subject matter.

[7] http://tools.aximag.fr/macau/chamilo-macau/.

As stated in the introduction, Google Translate does not translate well domain-specific terms. Our students solved this problem by creating and "feeding" a specific online multilingual terminological lexicon in a Google Spreadsheets file. This allowed them to maintain inter-translator consistency in their translations.

Table 1. Current status of the MACAU-Chamilo platform

Subject matter	Content type	Pages (html)	Available translations
Introduction to propositional and first-order logic	Full book	45	Chinese (full) English (partial) Russian (partial)
Computational complexity	Lecture notes	13	Chinese (full)
Human-machine interaction	Teacher lectures	7	Chinese (full)
Formal languages and parsing	Teacher lectures, hand-outs	5	Russian (partial)
Modelling of digital systems	Exam paper	2	Chinese (full)
AI and automatic planning	Exam paper	2	Chinese (full)
Introduction to ergonomics	Student report	1	Chinese (full)

As SECTra associates a chronometer to each segment, which measures the *primary PE time* (`Tpe_1`) spent in editing the PE cell of a segment, it is easy to retrieve it, for each segment post-edited in the SECTra PE interface. However, this is not yet possible directly for segments that have been post-edited from the iMAG palette, directly on a Web page. But a recent study indicates that `Tpe_1` can be computed from the combined edit distance[8] between the MT result used and the post-edited segment.

The *total PE time* (`Tpe_tot`) includes `Tpe_1` and the *secondary PE time* (`Tpe_2`), which is the time spent "outside a PE text area" (palette or SECTra cell), in particular on terminological search. In our experiments, we asked the participants to time their time globally, from the start to the end of a PE session. We then obtained a global value for the final ratio `Tpe_tot/Tpe_1` (about 3), from which, assuming proportionality, we got an estimated value for `Tpe_2` and `Tpe_tot` for each segment.

To estimate the gains in overall time, we must compare with the time it would take junior translators to produce comparable final translations in the classical way. The times reported in the profession (by technical translation agencies) are: (1) 1 h/p[9] to produce a first draft and (2) 20 mn/p to do a professional revision (by a senior translator). As what we achieve (and want to achieve) with our method is only the equivalent of a first draft, our basis of comparison is 1 h/p. As `Tpe_tot` is equal to 15–20 mn/p, we can

[8] For 2 strings A, B and 2 words u, v, $D_{comb}(A, B) = \alpha\, D_{char}(A, B) + (1-\alpha)\, D_{words}(A, B)$, where $Cost\_exchange(u, v) = D_{char}(u, v)$. This kind of mixed TER has ben introduced in 2004.

[9] h/p = hour per page, mn/p = minute per page, where a standard page has 250 words.

conclude that, in our context, post-editing machine translations is three to four times faster than translating from scratch.

5 Conclusion and Perspectives

We have presented an e-learning platform that allows users to access educational content in many languages, and the method for producing multilingual content from monolingual sources. We have made a set of documents freely accessible on our platform[10] in Chinese (also some in Russian), achieving a satisfactory linguistic quality and a very good usage quality through the contributions of foreign students themselves.

Our perspectives are now:

- to increase the number of subject matters
- to recruit more post-editors and foster international collaborations, and to orchestrate their post-editing activities via the collaboration platform Synaps[11]
- to integrate some lexical and terminological helps, to ensure terminological coherence
- to make it possible to also easily edit the source segments (source post-editing!), to check and modify on screen the segmentation graph (at segment level), and to control the normalization results
- to use the obtained HQ translation memories to train custom statistical machine translation systems — that has already been done for French→Chinese [5]

References

1. Green, S., Heer, J., Manning, C.D.: The efficacy of human post-editing for language translation. In: Proceedings of ACM Human Factors in Computing Systems (2013)
2. Tarasowa, D., Khalili, A., Auer, S., Unbehauen, J.: CrowdLearn: crowd-sourcing the creation of highly-structured e-learning content. In: 5th International Conference on Computer-Supported Education CSEDU 2013 (2013)
3. Boitet, C., Phap, H.C., Nguyen, H.T., Bellynck, V.: The iMAG concept: multilingual access gateway to an elected web sites with incremental quality increase through collaborative post-edition of MT pretranslations. In: TALN-2010, 8 p (2010)
4. Besacier, L.: Traduction automatisée d'une œuvre littéraire: une étude pilote. Traitement Automatique du Langage Naturel (TALN), juillet 2014, Marseille, France (2014)
5. Wang, L., Boitet, C.: Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. In: Proceedings of MT Summit XIV, the 2nd Workshop on Post-editing Technologies and Practice (2013)

[10] http://tools.aximag.fr/macau/chamilo-macau/.
[11] https://synaps.me/.