Multilingual Access to Educational Material through Contributive Post-editing of MT Pretranslations by Foreign Students

Ruslan Kalitvianski, Valérie Bellynck, Christian Boitet

Multilingual Access to Educational Material through Contributive Post-editing of MT Pretranslations by Foreign Students

Ruslan Kalitvianski*,**, Valérie Bellynck* Christian Boitet*

LIG-GETALP, bâtiment IMAG, 700 avenue Centrale, 30841 Grenoble cedex 9
prénom.nom@imag.fr
*Viseo Technologies, 4 avenue du Doyen Louis Weil, 38000 Grenoble

Abstract. In our teaching practice, we often observe that due to the lack of prerequisites and limited mastery of a language, foreign students face difficulty 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 producing educational content in the native tongues of the students using monolingual sources, through a contributive computer-assisted multilingualization by voluntary participants. The process consists in post-editing MT (Machine Translation) pretranslations via an interactive multilingual access gateway, 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 (125K words) of French undergraduate and graduate course material about computer science, at a rate of about 10 minutes per standard page. This multilingual resource is freely accessible on the XXXXX-Chamilo platform.

Keywords: Multilingual access, 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].

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

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

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

adfa, p. 1, 2011.

- 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 their host university.

Motivated by this observation, we started our XXXX 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]. This service 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.
2. While Google Translate allows to suggest 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. Google Translate 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. Thus, in this paper, we propose to use machine translation as an aid to producing multilingual documents. We describe our approach to multilingual access via interactive Multilingual Access Gateways (iMAG) embedded into an e-learning platform, as well as the resources we were able to produce this way.

The rest of the paper is organized as follows: in the next section, we discuss previous and similar work, then we describe the platform and its features, and lastly we discuss the resources that have been produced, as well as the encountered difficulties.

2 Prior work

We present here work that implements some of the ideas that we propose for multilingual access to educational content.

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

2.1 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 programmes 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 lasting web service.

The translation tools were to be specifically adapted to translation of course syllabi. In 2013, we have evaluated the online Bologna demonstrator and found the translations to be of a quality inferior to that of Google Translate. This service has since been discontinued.

Several ideas implemented by the Bologna converge with those of our project:

- a 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 of different document formats (html, docx, xslx, txt, rtf, URL link)…

However, Bologna had both conceptual and implementation flaws.

- One can criticize the lack of ambition of the project, limited to translating 9 of 22 European languages into English and Chinese. International students arriving in a foreign country may not have a sufficient grasp of English to understand translated documents or to contribute to the improvement of machine pretranslations.
- 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.
- The access to the post-edition interface is restricted to approved users, and the interface itself is cumbersome. In order to elicit contributions, the interface should allow for post-edition directly on the displayed document, freely accessible to the public.

2.2 SlideWiki

SlideWiki is a recent project for online collaborative construction of educational presentations [1]. These presentations can either be built on the website, or imported from a pptx format.

An interesting aspect is the possibility of producing versions in a different language via a translation by Google. However, limiting the content type to presentations

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

appears too restrictive, and the lack of a translation memory limits the efficiency of the translation process.

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 [2].

An iMAG is a gateway very much like Google Translate at first sight: one specifies a 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 bubble displays the source segment and proposes to contribute by correcting the target segment, in effect post-editing an MT result.

Contrary to Google Translate, 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 pretranslated 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. The post-editing bubble 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 to select, parameterize and call the MT systems and 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 corpus and MT manager, SECTra_w.

MT systems tailored to the selected sublanguage can be built and have ben built (by combinations of empirical and expert methods) from the TM dedicated to a given elected Web site. That approach will inherently raise the linguistic and terminological quality of the MT results, hopefully converting them from rough into raw translations.

Fig. 2. The advanced post-editing interface

Besacier [3] reported on an experiment on collaborative translation into French of 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 platfotrm 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 document and the 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 translation should be implemented.

We therefore propose to use iMAGs for multilingual access. For our course repository, we chose to use the open-source e-learning platform1 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 and define terminological dictionaries.

We equipped a Chamilo platform with tools for selecting the language of access of a document. The list of languages is defined by the available MT systems or translation memories.

Multilingual access integrated into Chamilo.

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 post-editions, which are both stored in a TM (translation memory) managed by the gateway. This works for any HTML content on 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 courses.

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 appear 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 score that the user assigns to the translation is used for ranking the translations and post-editions in the memory; the one with the highest score is displayed during the visit of the page via the iMAG.

This 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 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 only displays 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 are post-edited but await moderation (for contributions by users that are not registered), and red are for raw machine translation results.

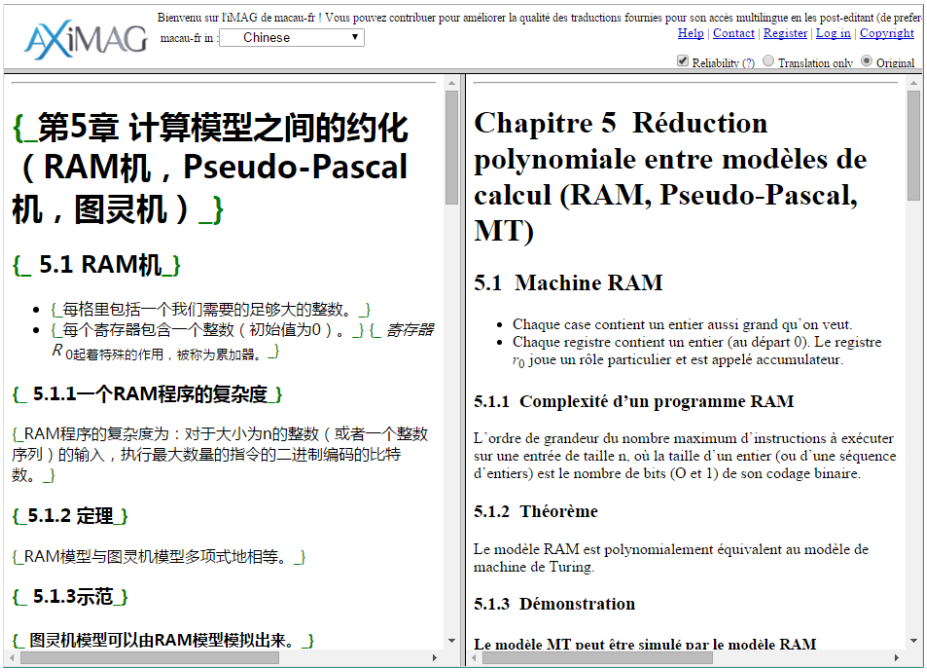

Fig. 3. A parallel presentation of a translation and the source document, displaying optional reliability brackets.

4 Results

We want to confirm the hypothesis that post-editing machine translations of course material by volunteers is a viable way of producing multilingual documents, even when the machine translations themselves may not be of good quality.

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 X. XXXX, Y. YYYY and Z. ZZZZ), lecture notes on computational complexity, as well as various hand-outs.

Documents came in different formats. The book and lecture notes were in LaTeX and had to be converted into HTML via conversion tools such as HeVeA[Footnote] and LaTeX2HTM[Footnote]. Others were in Microsoft Word's DOCX format, an XML-based format whose conversion into HTML is straightforward and well performed by office suites such as Microsoft Office, LibreOffice and Abiword.

(At the time when this experiment was conducted, there were no tools for converting PDF files into HTML documents of acceptable quality. The available tools either only extracted the text, disregarding document layout, or produced HTML documents that attempted to preserve the typographical layout of the pages, but in doing so produced HTML which was difficult to parse for MT systems. Progress has since been made by Microsoft Word, which now allows to transform some PDF files into Word documents.)

These documents then had to be segmented into pages of convenient size for the MT system we used (Google Translate 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 alphanumeric 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 has 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 (thus over 500 standard pages of 250 words) have been post-edited into Chinese.

In table 1 we report …

Table 1. Current status of our platform

Subject	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 stated in the introduction, Google Translate does not translate well domain-specific terminologies. Our students resolved this problem by constituting a lexicon in a Google Spreadsheets file. This allowed them to maintain inter-translator consistency in their translations.

We measured time spent on segments … (Il faut dire que c'est effectivement plus rapide que de traduire de zéro. On sait qu'on passe de 1h à 15-20 minutes par page standard, mais quelqu'un a une référence pour cela ?). On peut aussi mentionner l'article controversé de Garcia, qui montre que … (voir l'article de de Laurent sur Powers)

In his experiment, Garcia [4] reported that post-editing….

Combien de segments ont-ils post-édité ? d'après un mail de CB du 19/11/2014, « Nos thésards chinois et des étudiants chinois de M1-info ont post-édité vers le chinois environ 30000 « segments » (phrases ou titres), soit près de 1500 pages « standard » (de 250 mots) : articles et annonces scientifiques, supports et notes de cours. », mais quelle est la part MACAU ?

A thing worth noting is that it proved to be quite useful for some as it helped them prepare some exams requiring a good and fast understanding of quite lengthy and complex exam papers.

All these documents are freely accessible to anyone, via a link which will be provided in the camera-ready version.

- shown that our approach to multilingualisation "works"
- produced a significant amount of documents in Chinese, with free access
- seen that post-editing helps understand the subject matter (ex: Wu Jiang).

There are several difficulties associated with the process described above:

- Documents come in various formats, some of them being hard to transform automatically into HTML. This mainly concerns pdfs and some LaTeX files. Teachers are not always able to provide their source documents, only pdfs.
- Segmentation is a complex task and has to be done at two levels, and intertwined with normalization.
- Normalization is also delicate. It consists in removing or protecting from translation formulas, entities, non-linguistic elements (icons, etc.), and formatting tags. It should also "derecusivize" recursive pieces of text (such as footnotes or values of TITLE attributs in an html anchor).

Segmentation and normalization should be performed at two levels: first at a *general level* (hierarchical structuring, segment separation, derecursivization), and second at a *specific level*, for each particular MT system to be called. For example, Systran should be called using its "XML flow", with translation units larger than sentences (at least paragraphs), while Moses-based systems should be called segment by segment.

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 showed that this method is efficient, and have produced a set of documents that are freely available on our platform.

The perspectives are:

- to increase the number of subject matters
- to recruit more post-editors and foster international collaborations
- 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 normalisation results.

6 References

1. Tarasowa, D., Khalili, A., Auer, S. & Unbehauen, J. (2013) CrowdLearn: Crowd-sourcing the Creation of Highly-structured E-Learning Content. 5th International Conference on Computer-Supported Education CSEDU 2013.
2. Boitet, C., Phap, H. C., Nguyen, H. T. & Bellynck, V. (2010) The iMAG concept: multilingual access gateway to an elected Web sites with incremental quality increase through collaborative post-edition of MT pretranslations. TALN-2010, 8 p.
3. Besacier, L. (2014) Traduction automatisée d'une œuvre littéraire: une étude pilote. Traitement Automatique du Langage Naturel (TALN), juillet 2014, Marseille, France.

4. Garcia, I.: Translating by post-editing: is it the way forward? Journal of Machine Translation, 25(3), 217–237. (2011)