Demo of iMAG possibilities: MT-postediting, translation quality evaluation, parallel corpus production

WANG Ling Xiao, ZHANG Ying, Christian BOITET, Valerie BELLYNCK
Lingxiao.Wang@imag.fr, Ying.Zhang@imag.fr,
Christian.Boitet@imag.fr, Valerie.Bellynck@imag.fr

ABSTRACT

An interactive Multilingual Access Gateway (iMAG) dedicated to a web site S (iMAG-S) is a good tool to make S accessible in many languages immediately and without editorial responsibility. Visitors of S as well as paid or unpaid post-editors and moderators contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. Pre-translations are produced by one or more free MT systems. Continuous use since 2008 on many web sites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less. There are two interesting side effects obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora and to set up a permanent task-based evaluation of multiple MT systems. We will demonstrate (1) the multilingual access to a web site, with online postediting of MT results "à la Google", (2) postediting in "advanced mode", using SECTra_w as a back-end, enabling online comparison of MT systems, (3) task-oriented built-in evaluation (postediting time), and (4) application to a large web site to get a trilingual parallel corpus where each segment has a reliability level and a quality score.

KEYWORDS: Online post-editing, interactive multilingual access gateway, free MT evaluation

TITLE AND ABSTRACT IN CHINESE

iMAG功能展示：
机器翻译后编辑，翻译质量评估，平行语料生成

简述

一个iMAG (interactive Multilingual Access Gateway, 多语言交互式网关) 是很好的面向一个网站的工具，它可以提供对该网站的多语言访问，并且无需任何编辑。通过iMAG访问该网站的用户，可以作为有偿或无偿的后编辑人员或是管理者，来对该网站的文本段进行可持续的、增量的改进。该网站的预翻译是由一个或多个免费的MT系统提供的。自从2008年以来，通过iMAG对多个网站进行多语言的持续访问结果表明，对于相对翻译质量，首轮由初级翻译者提供的翻译，使用iMAG只占纯人工翻译40%的时间，或更少。iMAG有两个非常吸引人的方面并且无需额外成本：iMAG能用于产生高质量的平行语料，而且可以通过多个MT系统对其进行长久性的评估。我们将要展示：(1) 多语言访问目标网站，并对Google提供的预翻译进行在线后编辑，(2) 后编辑的高级模式，SECTra作为后台模块，可实现MT系统的在线比较，(3) 面向任务的评估 （后编辑时间），和 (4) 应用到大型网站，可获得三种语言的平行语料，每个文字段都拥有可靠性和质量的评分。

关键词：在线后编辑，多语言交互网关，免费MT评估

Proceedings of COLING 2012: Demonstration Papers, pages 475–482,
COLING 2012, Mumbai, December 2012.

1 Introduction

An iMAG is a website used as a gateway allowing a multilingual access to one (in general) or several elected websites. The name "iMAG" stands for interactive Multilingual Access Gateway.

Apparently, an iMAG is similar to existing well-known translation gateways such as Google Translate, Systran, Reverso, etc. The first essential difference is that an iMAG is only used for elected websites. This allows the iMAG to manage the multilingualization of certain websites better than existing translation gateways. With an iMAG, we can enhance the quality of translated pages, starting from raw output of general-purpose and free MT servers, usually of low quality and often understandable unless one understands enough of the source language.

An iMAG is dedicated to an elected website, or rather to the elected sublanguage defined by one or more URLs and their textual content. It contains a translation memory (TM), both dedicated to the elected sublanguage. Segments are pre-translated not by a unique MT system, but by a (selectable) set of MT systems. Systran and Google are mainly used now, but specialized systems developed from the post-edit part of the TM, and based on Moses, will be also used in the future.

iMAG also contains a module SECTra (Système d'Exploitation de Corpus de Traductions sur le web), in English, "Contributive Operating System of Translation Corpora on the Web". SECTra is a Web-based system offering several services, such as supporting MT evaluation campaigns and online post-editing of MT results, to produce reference translations adapted to classical MT systems not built by machine learning from a parallel corpus.

2 Manipulation and pre-translation in the iMAG page

2.1 Access multilingual website by iMAG

Figure 1 shows the iMAG access interface to LIG (the Grenoble computer science laboratory) website. We choose target language (Chinese) in the pull-down menu. The page is now accessed in Chinese language. One or more free MT servers, in this case Google Translate and Systran, produce initial translations.

Figure 1: Access website of Grenoble computer science laboratory by iMAG

2.2 Post-edition and evaluation in web page context

In iMAG, user can also optimize the translation results. As shown in Figure 2, when the user moves the mouse on translation unit (for example: a word, a title), the system will automatically pop up a small dialog box. This dialog box display source language content in blue font, and user can post edit and evaluate the translation results.

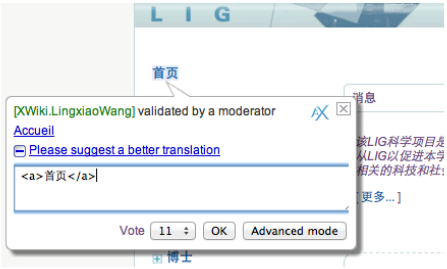

Figure 2. Optimize translation results in iMAG interface

If user is an anonymous, or non-privileged, this optimize translation and ratings only display to him, and he can't enter the advanced mode. If user has privilege, optimize translation and ratings will be stored in the system database, and also display to publics. If database contains multiple optimizes translations, system will select translation, which has the highest scores and time recently. For those users who have the appropriate permissions, they can come into "Advanced mode", and arrives into SECTra. This will be described in chapter 3.

2.3 Visualization of the translation quality

In the translated page, users can view quickly and clearly the translation quality of web pages by “reliability” mode. As shown in Figure 3, a color bracket encloses each translation unit. If user post-edit in this page, then his result will be displayed directly on the page, at the same time, bracket's color will be changed based on user permissions. Green brackets indicate that the translation results are edited and saved by privileged user. Orange means the translation results are edited and saved locally by anonymous users (only for anonymous users). Red indicate the translation results have never been edited. If the user clicks on the Original button, the left side of the browser will display the translation results; the right side displays the source language page.

Figure 3. iMAG page display in “reliability”, “original” mode

2.4 Interaction between iMAG and SECTra

Another possibility is to use the advanced mode (see the chapter 2.2), which consists in post-editing a pseudo-document that is in fact a part of the translation memory.

Remark: content of chapter 2 will be on display in the video 1.

3 Post-edition of TMs in "Advance Mode" (SECTra)

In order to obtain the translation results with the high quality, post-editing is a very important point, but also the most time-consuming work. In "Advance Mode" (SECTra), we can quickly get high quality translation results with minimum price. Figure 4 shows the interface of SECTra post-editing features.

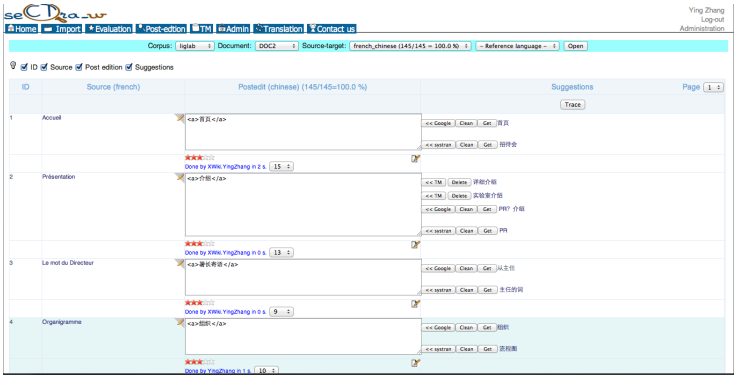

Figure 4: Advanced mode (SECTra screen).

3.1 Proposition of translation in SECTra

The first time user create an iMAG, he can select different machines translations systems for his website. Certainly he can also add new machine translation system later in SECTra. In interface of post edit, SECTra allows us to do operations for machine translation results (such as Google Translate, Systran translation), and translation memory database.

- For machine translation: clear translation result, re-call the machine translation system, and use the translation result.
- For translation memory: delete translation memory, use translation memory

3.2 Comparison between current translation and MT/TM results

As shown in Figure 5, users can compare distance between the current translation and translation memory, or between the current translation and machine translation.

Figure 5: Comparison between current translation and MT/TM results

SECTra can also provide a reference language, which helps users to better post-edit, as shown in Figure 6.

Figure 6: Interface with reference language

3.3 Post-edition and evaluation

Users can also vote the number of stars and the value of rating for these post-editions. The number of stars is the control ability of the language pair (the source language and the target

language) of the current post-edition. The value of rating is the satisfaction level of the current post-edition.

In the process of post-edition, the system will automatically record the time and segments number. As the first two authors are Chinese, they have experimented with French-Chinese (on a town web site) and with Chinese-English (on a Chinese web site dedicated to a famous Chinese NLP scientist). Here are the results.

Language pair	Human PE time	Human first draft time	# segments	# source words (or characters)	# target words (or characters)
Fr->Zh	17 mins	72 mins	76	303 (Fr) — 1.16 p.	519 (Zh) — 1.3 p.
Zh->En	16 mins	75 mins	32	495 (Zh) — 1.25 p.	307 (En) — 1.16 p.

3.4 Visualization of post-edition in iMAG Web pages

Post-edition results will be displayed directly on the iMAG Web page, and bracket's color will be changed based on user permissions (see the chapter 2.3).

Remark: content of chapter 3 will be on display in the video 2.

4. To obtain a high-quality parallel corpus

4.1 Filtering and selection of segments

On the platform of the SECTra, the user can export corpus of TM. At the time of export, we can filter segments by stars and scores. In Figure 7, for example the source language is French, the target language is Chinese, and we can select part of segments for export.

Figure 7. Interface of export corpus

4.2 Production of parallel corpus and download

For the selected parallel corpus, the system will generate two txt files, and users may download these files, the results shown in Figure 9.

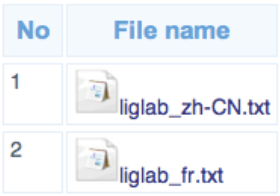

Figure 8. Corpus files

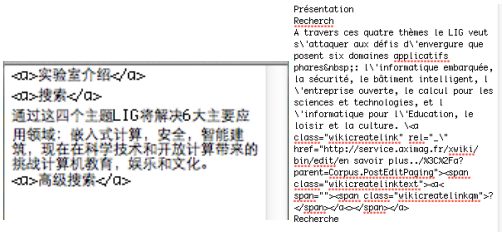

Figure 9. Downloaded corpus files

Remark: content of chapter 4 will be on display in the video 3.

5. Conclusion and perspectives

Continuous use since 2008 on many web sites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less.

In the near future, the system will be integrated Moses, and based on Moses for provide more accurate TA results.

References

(2009) *A Web-oriented System to Manage the Translation of an Online Encyclopedia Using Classical MT and Deconversion from UNL.* Proc. CCC 2009

(2010) *The iMAG concept: multilingual access gateway to an elected Web site with incremental quality increase through collaborative post-edition of MT pretranslations.*

(2008) *an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora.* Proc. LREC-08, Marrakech, 27-31/5/08, ELRA/ELDA, ed., 8 p.

Huynh C.-P., Blanchon H. & Nguyen H.-T. (2008) *A Web-oriented System to Manage the Translation of an Online Encyclopedia Using Classical MT and Deconversion from UNL.* Proc. CI-2008 (WS of ASWC-08), Bangkok, 9/12/08, ACL, ed., 8 p.